Tight bound on relative entropy by entropy difference

David Reeb* and Michael M. Wolf†

Department of Mathematics, Technische Universität München, 85748 Garching, Germany

Abstract

We prove a lower bound on the relative entropy between two finite-dimensional states in terms of their entropy difference and the dimension of the underlying space. The inequality is tight in the sense that equality can be attained for any prescribed value of the entropy difference, both for quantum and classical systems. We outline implications for thermodynamics and information theory, such as a necessary condition for a process to be close to thermodynamic reversibility, or an easily computable lower bound on the classical channel capacity. Furthermore, we derive a tight upper bound, uniform for all states of a given dimension, on the variance of the surprisal, whose thermodynamic meaning is that of heat capacity.

Contents

*david.reeb@tum.de
†m.wolf@tum.de

1 Introduction

The relative entropy is a distance-like measure that appears in a multitude of areas, such as information theory, thermodynamics, statistics and learning theory, being of operational significance in various situations (see Section 3 for a few applications). Also known as the Kullback-Leibler divergence, it was first introduced for probability distributions [KL51], and later generalized to quantum states [Ume62]. Another ubiquitous quantity is the entropy of a probability distribution or quantum state [Sha48, vN32], which in Thermodynamics had already played a central role because entropy differences characterize possible and impossible thermodynamic state transformations (see e.g. the Clausius inequality in Section 3.1.1).

In this work, we lower-bound the relative entropy $D(\sigma\|\rho)$ between two states σ, ρ (probability distributions or quantum states) in terms of their entropy difference $\Delta = S(\sigma) - S(\rho)$. Qualitatively, it is clear that such non-trivial lower bounds exist in any finite dimension due to the compactness of the state space, since $\Delta \neq 0$ implies $\sigma \neq \rho$ and thus $D(\sigma\|\rho) > 0$ by Klein's inequality [OP93]. Our main inequality (Theorem 1) makes this quantitative and is furthermore *tight*, meaning that for each dimension d it provides the best lower bound on $D(\sigma\|\rho)$ in terms of Δ, both for classical and quantum systems.

We note that any lower bound that can be derived by combining the tight Pinsker inequality [Csi67, AE05] with the tight Fannes-Audenaert inequality [Fan73, Aud07] will *not* be tight and will be strictly weaker than the derived bounds, even in its functional dependence (Remark 6).

Also considering states of finite dimension d in Section 2.2, we give a tight upper bound on the variance of the *surprisal* (or information gain), which is quadratic in $\log d$ (Section 2.2.1); of course, the expectation value of the surprisal is just the entropy and is bounded by $\log d$ [OP93]. One thermodynamic implication of this result is an upper bound on the heat capacity of finite-dimensional systems (Section 2.2.2).

The inequalities presented here arose out of, and are used in, an investigation of finite-size effects in Landauer's Principle [RW13], but we expect them to have applications elsewhere in thermodynamics and information theory; some are outlined in Section 3. Furthermore, the finite-size bounds here arise in one-partite systems, whereas the Landauer scenario – the topic of [RW13] – is bipartite, involving a system and a thermal reservoir [Lan61].

Physically, our bounds are especially interesting for quantum thermodynamics [GMM10, SBL$^+$11] and generally for the thermodynamics of microscopic systems or devices. Furthermore, even a large heat bath may sometimes be reasonably treated as small, when the equilibration time with another system is so short that only a small part of the bath effectively interacts with the system. Our bounds can be applied to derive finite-size corrections to well-known physical laws and, for example, alter efficiency analyses of physical process like Carnot's or Landauer's [AG12, RW13].

By treating the Shannon and von Neumann (relative) entropies, our results are relevant to the conventional situation of many independent copies of a system state ("thermodynamic limit"), averaging quantities over these copies ("ensemble averages"). Thermodynamics and information theory can instead also be examined in the "single-shot setting", necessitating extra parameters such as the success probability of a process (e.g. [Ren05, Abe11, EDR$^+$12]). Our setup is thus different from the one-shot scenario: Whereas the latter concerns a *finite (small) number* of systems, our results have implications in the limit of infinitely many *finite-dimensional* system copies. The variance computed in Section 2.2.1, however, can quantify how many copies of a finite-dimensional system have to be averaged before the Shannon or von Neumann entropies become sensible measures (see also [TH12, Li12]).

1.1 Notation

All states σ, ρ will be on a space of finite dimension $d < \infty$. In the quantum framework, states are positive semi-definite $d \times d$-matrices of trace 1 ("density matrices" [NC00]); in the classical (probability theory) framework, they are probability distributions on d atomic events [CT06] and can be identified with diagonal density matrices in the obvious way. We often require $d \geq 2$ to exclude the trivial one-dimensional case, where some statements become pathological.

The *entropy* of a state ρ is defined as

$$S(\rho) \;:=\; -\operatorname{tr}\left[\rho \log \rho\right] \;. \tag{1}$$

Throughout, we use the natural logarithm, denoted by log, and employ the usual rules of calculus on the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, such as $0 \log 0 := 0$; only in Section 3.2.1 will we also use the D-ary logarithm $\log_D x := (\log x)/(\log D)$, with $D > 1$. A quantity of central interest will be the *entropy difference* $\Delta \equiv \Delta(\sigma, \rho)$ of the states σ and ρ:

$$\Delta(\sigma, \rho) \;:=\; S(\sigma) - S(\rho) \;\in [-\log d, +\log d] \;. \tag{2}$$

The other central quantity is the *relative entropy* between two states σ and ρ:

$$D(\sigma \| \rho) \;:=\; \operatorname{tr}\left[\sigma \log \sigma\right] - \operatorname{tr}\left[\sigma \log \rho\right] \;, \tag{3}$$

which equals $+\infty$ if $\operatorname{supp}[\sigma] \not\subseteq \operatorname{supp}[\rho]$, and is finite otherwise, non-negative, and vanishes iff $\sigma = \rho$.

We also define binary versions of the entropy and relative entropy, i.e. for binary probability distributions $(x, 1-x)$ and $(y, 1-y)$ with $0 \leq x, y \leq 1$:

$$H(x) \;:=\; S\left(\operatorname{diag}(x, 1-x)\right) \qquad = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x} \;, \tag{4}$$

$$D_2(x \| y) \;:=\; D\left(\operatorname{diag}(x, 1-x) \| \operatorname{diag}(y, 1-y)\right) \qquad = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} \;. \tag{5}$$

Note that the entropy difference $\Delta(\sigma, \rho)$ changes sign under exchange of σ and ρ, whereas the relative entropy $D(\sigma \| \rho)$ does not generally have any symmetry under exchange. For example, $\Delta = -\log d$ forces ρ to be the maximally mixed state $\mathbb{1}/d$ and σ to be any pure state (any Hermitian projector of rank 1), resulting in $D(\sigma \| \rho) = \log d$; whereas $\Delta = +\log d$ interchanges these ρ and σ and gives $D(\sigma \| \rho) = \infty$. The latter case is special as for any other $\Delta \in [-\log d, \log d)$ there exist full-rank states σ and ρ with $\Delta(\sigma, \rho) = \Delta$, such that $D(\sigma \| \rho) < \infty$ is finite.

For a more detailed discussion of entropic quantities we refer to [OP93] and [Weh78] or, in the context of classical and quantum information theory, to [CT06] and [NC00].

The acronyms LHS and RHS mean "left-hand side" and "right-hand side", respectively.

2 Main results

In Section 2.1 we state the tight inequality between relative entropy and entropy difference (Theorem 1) and describe properties and simplifications of the bound that are useful for applications (Section 3). The tight upper bound on the variance of the surprisal (or heat capacity) is given in Section 2.2. The proofs follow in Section 4.

2.1 Relative entropy vs. entropy difference

To state our main inequality and its simplifications, we define for $d \geq 2$ and $\Delta \in [-\log d, \log d]$:

$$M(\Delta, d) \;:=\; \min_{0 \leq s,r \leq (d-1)/d} \left\{ D_2(s\|r) \;\middle|\; H(s) - H(r) + (s-r)\log(d-1) = \Delta \right\} \;, \tag{6}$$

$$N(d) \;:=\; \max_{0<r<1/2} r(1-r) \left(\log \frac{1-r}{r}(d-1) \right)^2 \;, \tag{7}$$

$$N_d \;:=\; \frac{1}{4}\log^2(d-1) + 1 \;. \tag{8}$$

(The expression $\log^2(d-1)$ should always be read as $(\log(d-1))^2$.) All of these quantities can be efficiently computed numerically as they involve optimizations over at most two bounded real variables. See also Fig. 1, and Lemmas 13 and 14 (Section 4.3) for relations among (6)–(8).

Theorem 1 (Tight lower bound on relative entropy by entropy difference)**.** *Let σ, ρ be states of dimension d, with $2 \leq d < \infty$, and define $\Delta := S(\sigma) - S(\rho)$. Then:*

$$D(\sigma\|\rho) \;\geq\; M(\Delta, d) \;, \tag{9}$$

with the function $M(\Delta, d)$ defined in Eq. (6).

Conversely, for any $\Delta \in [-\log d, \log d]$, there exist σ, ρ attaining equality in (9). More precisely, for any pair (s, r) attaining the minimum in (6), the commuting d-dimensional states

$$\sigma \;:=\; \operatorname{diag}\left(1-s, \frac{s}{d-1}, \ldots, \frac{s}{d-1}\right) \;, \quad \rho \;:=\; \operatorname{diag}\left(1-r, \frac{r}{d-1}, \ldots, \frac{r}{d-1}\right) \tag{10}$$

have entropy difference $S(\sigma) - S(\rho) = \Delta$ and achieve equality $D(\sigma\|\rho) = M(\Delta, d)$.

Theorem 2 (Properties of the tight bound)**.** *Let $2 \leq d < \infty$. Then the function $M(\Delta, d)$ in the tight lower bound (9) is non-negative, continuous and strictly convex in $\Delta \in [-\log d, \log d]$, and continuously differentiable in the interior of this interval. It takes values $M(0, d) = 0$, $M(-\log d, d) = \log d$, $M(\log d, d) = \infty$, and $M(\Delta, d) < \infty$ for $\Delta \in [-\log d, \log d)$.*

For any $N \geq N(d)$, with $N(d)$ from Eq. (7), the following lower bounds hold for all Δ:

$$M(\Delta, d) \;\geq\; N\left(e^{\frac{\Delta}{N}} - 1 - \frac{\Delta}{N}\right) \;\geq\; \frac{\Delta^2}{2N} + \frac{\Delta^3}{6N^2} \;, \tag{11}$$

$$M(\Delta, d) \;\geq\; \frac{\Delta^2}{3\log^2 d} \;. \tag{12}$$

Easily computable choices for N are $N = N_d = \frac{1}{4}\log^2(d-1) + 1 > N(d)$, or $N = \log^2 d > N(d)$.

Remark 3 (Equality cases in Eq. (9))**.** Regarding the equality statement in Theorem 1, we remark that for any $\Delta \in [-\log d, \log d]$ the minimum in (6) actually exists, i.e. is attained for some pair (s, r) (see Section 4.2), and equals $\infty \in \overline{\mathbb{R}}$ for $\Delta = \log d$. Note that for states of the form (10), it is $D(\sigma\|\rho) = D_2(s\|r)$, $S(\sigma) = H(s) + s\log(d-1)$ and similar for $S(\rho)$, which explains the connection between (9) and (6). In Remark 11 we elaborate on the states (10), which come all from the same exponential family.

For $\Delta \neq 0$, the pair (σ, ρ) from (10) constitutes, up to simultaneous unitary equivalence, the unique d-dimensional states achieving equality $D(\sigma\|\rho) = M(\Delta, d)$ and $S(\sigma) - S(\rho) = \Delta$. This follows from the proof of Theorem 1 in Section 4.1 as the optimal states for $\Delta \neq 0$ are necessarily

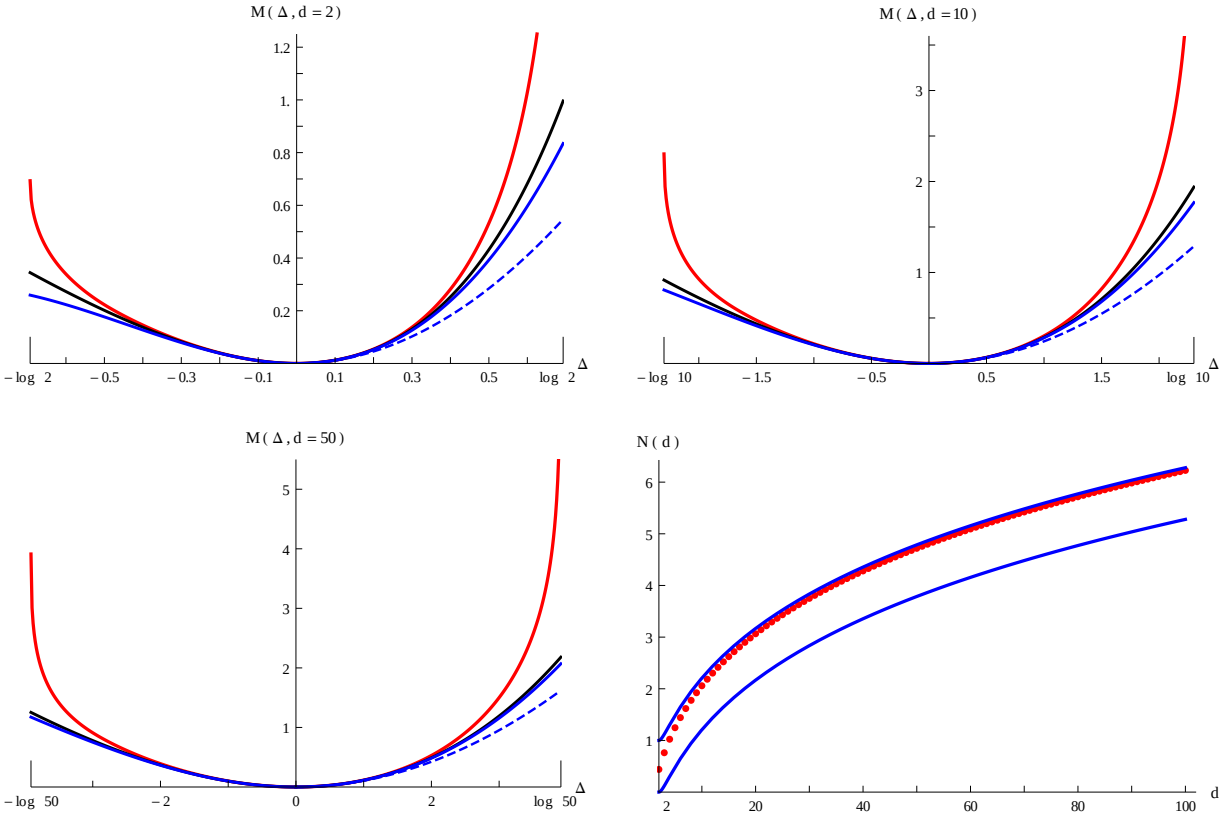

Figure 1: *Upper and left panels:* The upper (red) curves show $M(\Delta, d)$ from Eq. (6) (tight lower bound of Theorem 1) for $d = 2, 10, 50$. The black and blue solid curves below are the lower bounds from Eq. (11) with the optimal $N = N(d)$, the dotted blue curve is the quadratic lower bound $\Delta^2/2N(d)$ (for $\Delta \geq 0$). At $\Delta = \pm \log d$, all these lower bounds approach 2 in the limit $d \to \infty$, whereas $M(-\log d, d) = \log d$ and $M(\log d, d) = \infty$. *Lower right panel:* The red dots show $N(d)$ for $2 \leq d \leq 100$ (Eq. (7)), which approaches its easily computable upper bound N_d (Eq. (8) and Remark 9) as $d \to \infty$ and which is lower-bounded by $N_d - 1$ (blue curves).

of the form (10) with $0 \leq s, r \leq (d-1)/d$, and since for $\Delta \neq 0$ the pair (s, r) attaining the minimum in (6) is unique (which is shown in our proof of the convexity of $M(\Delta, d)$ in Section 4.2). For $\Delta = 0$, exactly the pairs with $\sigma = \rho$ attain equality in (9).

As inequality (9) is tight for *commuting* density matrices, it is tight for classical probability distributions (diagonal density matrices) as well.

Remark 4 (Goodness of the lower bounds in Eq. (11))**.** One can check that the function $M(\Delta, d)$ is smooth around $\Delta = 0$ (see Section 4.2) and that the RHS of (11) with $N = N(d)$ is its cubic Taylor expansion. This is thus the best cubic lower bound possible, and $\Delta^2/2N(d)$ is the best quadratic lower bound for $\Delta \geq 0$ (Fig. 1); it is however not a lower bound for small $\Delta < 0$.

The lower bounds in (11) are quite good (cf. Fig. 1) even for relatively large $|\Delta|$: For any constant $t \in (-1, 1)$, the states (10) with $s = (1+t)/2$, $r = (1-t)/2$ give $\Delta = \Delta(\sigma, \rho) = t \log(d-1)$ and $M(\Delta, d) \leq D(\sigma\|\rho) = D_2(s\|r) = t \log(1+t)/(1-t)$, whereas the lowest bound in (11) gives $\simeq 2t^2$ (the cubic term vanishes as $\sim 1/\log d$). Even for the large values $\Delta = \pm\frac{1}{2} \log d$, the lower bound is thus tight (for large d) up to at most 10%.

The quantity $N(d)$ from Theorem 2 appears in the upper bound in Theorem 8 as well, see Remark 9. For this quantity, see also the lower right panel in Fig. 1.

Remark 5 (Monotonicity of $M(\Delta, d)$ in Δ)**.** The tight lower bound $M(\Delta, d)$ is strictly monotonically decreasing for $\Delta \leq 0$ and strictly increasing for $\Delta \geq 0$ (cf. Fig. 1). This follows since the non-negative function $M(\Delta, d)$ vanishes at $\Delta = 0$ and is strictly convex by Theorem 2 (see also Fig. 1). As our convexity proof in Section 4.2 is quite involved, we give now a simpler proof of monotonicity. We actually prove

$$M(\lambda\Delta, d) \;<\; \lambda\, M(\Delta, d) \qquad \text{for } \Delta \in [-\log d, \log d] \setminus \{0\}\,,\ \lambda \in (0,1)\ . \tag{13}$$

Let first $\Delta \in (0, \log d)$, $\lambda \in (0,1)$, and let σ, ρ be states with $D(\sigma\|\rho) = M(\Delta, d)$ and $S(\sigma) - S(\rho) = \Delta$. Define states $\sigma_\mu := \mu\sigma + (1-\mu)\rho$ for $\mu \in [0,1]$. As $S(\sigma_\mu)$ is continuous in μ, there exists $\mu' \in (0,1)$ with $S(\sigma_{\mu'}) - S(\rho) = \lambda\Delta$, and by strict concavity of the entropy we have

$$\lambda\Delta \;>\; \mu' S(\sigma) + (1-\mu')S(\rho) - S(\rho) \;=\; \mu'\Delta\ , \tag{14}$$

i.e. $\mu' < \lambda$. Convexity of the relative entropy [OP93] finally gives

$$M(\lambda\Delta, d) \;\leq\; D(\sigma_{\mu'}\|\rho) \;\leq\; \mu' D(\sigma\|\rho) + (1-\mu')D(\rho\|\rho) \;<\; \lambda M(\Delta, d)\ . \tag{15}$$

(13) holds for $\Delta = \log d$ as well, since $M(\lambda\Delta, d) < \infty$ due to $\lambda\Delta < \log d$. The proof for $\Delta < 0$ is similar, now replacing ρ by some state $\rho_{\mu'} = \mu'\rho + (1-\mu')\sigma$.

Remark 6 (Lower bounds from the Fannes-Audenaert and Pinsker inequalities)**.** A weaker lower bound on the relative entropy $D(\sigma\|\rho)$ in terms of the entropy difference $\Delta = \Delta(\sigma, \rho)$, as in Theorem 1, can be obtained by combining the Fannes-Audenaert [Fan73, Aud07] and Pinsker [Csi67] inequalities: Writing $T := \|\sigma - \rho\|_1/2$ for the trace distance (or total variation or statistical distance [CT06]) between the states σ and ρ, we have the bound [Fan73, Aud07]

$$|\Delta| \;=\; |S(\sigma) - S(\rho)| \;\leq\; T\log(d-1) + H(T) \;=:\; h_d(T) \;\leq\; T\,(1 + \log(d-1) + \log 1/T)\ , \tag{16}$$

the first inequality being tight, and the sharpened Pinsker bound [Csi67, CT06, HOT81, AE05]

$$D(\sigma\|\rho) \;\geq\; s(T) \;\geq\; 2T^2\ , \tag{17}$$

where $s : [0,1] \to [0,\infty]$ is a function [AE05] such that the first inequality is tight (for any dimension $d \geq 2$) and which is lower-bounded by its quadratic Taylor expansion, $s(x) \geq 2x^2$.

If now $\Delta \in [-\log d, \log d]$ is given, we can invert the function $h_d|_{[0,(d-1)/d]} : [0, (d-1)/d] \to [0, \log d]$ from (16), or lower-bound the inversion of its RHS, to get a lower bound on T:

$$T \;\geq\; h_d^{-1}(|\Delta|) \;\geq\; \frac{e-1}{e}\,\frac{|\Delta|}{1 + \log(d-1) - \log|\Delta|}\ , \tag{18}$$

where the prefactor is $(e-1)/e \approx 0.63$. Plugging either of this into (17) yields a lower bound on $D(\sigma\|\rho)$. This approach, however, can never yield a quadratic lower bound $\sim \Delta^2$ near $\Delta = 0$, as (9) and (11)–(12) together do, since the tight lower bound $s(T)$ in (17) is quadratic near $T = 0$ and since h_d from (16) does *not* satisfy $h_d^{-1}(|\Delta|) \geq c(d)|\Delta|$ for any positive d-dependent constant $c(d)$. Numerically, one actually sees that, for all $d \geq 2$ and $\Delta \neq 0$, the lower bound obtained by plugging the RHS of (18) into the RHS of (17) is worse than the RHS of (11) with $N = N(d)$ (and even worse than the quadratic lower bound $\Delta^2/2N(d)$ for $\Delta > 0$).

Furthermore, this approach can only ever yield lower bounds that are invariant under $\Delta \mapsto -\Delta$ since the Fannes-Audenaert and Pinsker inequalities are both symmetric in σ and ρ. The tight lower bound $M(\Delta, d)$ however does actually not have this invariance (see Fig. 1).

Remark 7 (Dimension-independent bounds are trivial)**.** The non-trivial lower bounds (i.e., that are strictly positive for $\Delta \neq 0$) on the relative entropy from Theorems 1 and 2 depend explicitly on the dimension $d < \infty$. This has to be so as any dimension-independent bound will necessarily be trivial: Setting $t := \Delta/\log(d-1)$ in the states of Remark 4, with any constant $\Delta \in (-\infty, +\infty)$ and for large enough dimension d, gives $\Delta(\sigma, \rho) = \Delta$ and $D(\sigma\|\rho) = O\left(2\Delta^2/\log^2(d-1)\right) \to 0$ as $d \to \infty$, so that 0 is the best possible dimension-independent lower bound for any fixed value of Δ; this also holds for states over infinite-dimensional Hilbert spaces. In this case, however, the lower bound 0 is never attained for $\Delta \neq 0$ as $D(\sigma\|\rho) = 0$ would imply $\sigma = \rho$ [OP93, BR97], and thus $\Delta = 0$ (if the entropies $S(\sigma)$, $S(\rho)$ are at all defined).

We further remark that the optimal lower bound $M(\Delta, d)$ is a decreasing function of d, implying that the finite-size corrections in applications (see Section 3) will be smaller for larger systems. To see this, let $d' > d \geq 2$, $\Delta \in [-\log d, \log d]$, and let s, r be optimal variables when computing $M(\Delta, d)$ in (6). Now define $s' := s$, and find r' such that the entropy difference $\Delta(\sigma', \rho')$ between d'-dimensional states σ', ρ' as in (10) equals the given Δ; if $\Delta \neq 0$, r' will be closer to $s' = s$ than r is to s, such that $M(\Delta, d') \leq D_2(s'\|r') \leq D_2(s\|r) = M(\Delta, d)$ with strict inequality for $\Delta \neq 0$.

Proof of Theorems 1 and 2. The main part of the proof consists in reducing the minimization of $D(\sigma\|\rho)$ over (quantum) states σ, ρ with a fixed value of $\Delta(\sigma, \rho) = \Delta$ to the simpler minimization over two bounded real variables in (6). We give the full proofs in Sections 4.1–4.3. □

2.2 Dimension bounds on second moments

In Section 2.2.1 we derive a tight upper bound on the variance of the surprisal in terms of the dimension of the underlying space. Translating to thermodynamics in Section 2.2.2, this yields an upper bound on the second moment of the energy of thermal states or, equivalently, on the heat capacity of finite-dimensional systems.

The derived bounds have apparent connections to the relative entropy inequalities from Theorems 1 and 2. Namely, the optimal states are of the same form and the (optimal) bounds involve the same quantities (see Remarks 9 and 11). Also, all the bounds are dimension-dependent and become trivial for infinite-dimensional spaces (cf. Remark 7). Furthermore, the heat capacity bound of Corollary 10 is used in [RW13] in a bipartite situation to actually lower-bound a relative entropy term in an indirect way, as the direct bound by Theorem 1 would necessarily depend on an undesired entropic quantity, i.e. one of the "wrong" subsystem.

2.2.1 Maximum variance of the surprisal

In a classical random experiment described by a probability distribution $\rho = \mathrm{diag}(p_1, p_2, \ldots, p_d)$, the *information gain* upon outcome i is $(-\log p_i)$, which is the unique sensible information measure in the limit of many indepedent experiments [Sha48, CT06]. Equivalently, the surprise about obtaining i may be quantified by the *surprisal* $(-\log p_i)$. The (Shannon) entropy (1) is the expectation value of the surprisal, $S(\rho) = \sum_i p_i(-\log p_i) = \langle -\log\rho\rangle_\rho$. In this section, we look at its second moment, i.e. the *variance* or fluctuation of the surprisal:

$$\mathrm{var}_\rho(-\log\rho) \;:=\; \sum_i p_i(-\log p_i)^2 - \left(\sum_i p_i(-\log p_i)\right)^2 \;=\; \mathrm{tr}\left[\rho\left(-\log\rho - S(\rho)\right)^2\right]. \qquad (19)$$

In classical coding theory, when the source signals are i.i.d. distributed according to ρ, optimal prefix codes assign a codeword length of roughly $\simeq (-\log p_i)$ to symbol i [CT06]. The expected codeword length is thus $\simeq S(\rho)$ with fluctuation $\simeq \sqrt{\mathrm{var}_\rho(-\log\rho)}$, which implies a certain

fluctuation in the lengths of encoded messages. (This holds up to an overall factor logarithmic in the size of the code alphabet, see Section 3.2.1.) Similar second-order effects in hypothesis testing using only finitely many copies have recently been investigated in [TH12, Li12].

The above definitions in terms of a density matrix ρ are sensible in the quantum framework as well, and have similar interpretations [Sch95, NC00, SW01]. Note that $S(\rho)$ and $\mathrm{var}_\rho(-\log\rho)$ depend both only on the eigenvalues $\{p_i\}$ of the density matrix ρ.

Our main theorem here places a tight upper bound on the variance of the surprisal, only in terms of the dimension d of the system. A non-tight upper bound is implicit in [PPV10], where the term $\sum_i p_i \log^2 p_i$ in (19) has been bounded. For the expectation value of the surprisal, i.e. the entropy, a tight upper bound is of course well-known: $S(\rho) \leq \log d$.

Theorem 8 (Maximum variance of the surprisal)**.** *Let ρ be a state on a d-dimensional system. Then, for $d \geq 2$,*

$$\mathrm{var}_\rho(-\log\rho) \;\leq\; N(d) \qquad \equiv \max_{0<r<1/2} r(1-r)\left(\log\frac{1-r}{r}(d-1)\right)^2 \tag{20}$$

$$<\; N_d \qquad \equiv \frac{1}{4}\log^2(d-1)+1\;. \tag{21}$$

(Cf. definitions (7) and (8), and Lemma 14.) For $d = 1$, $\mathrm{var}_\rho(-\log\rho) = 0$.

For $d \geq 2$, let $r = r_d$ attain the maximum in (20). Then the d-dimensional state

$$\rho \;:=\; \mathrm{diag}\left(1-r_d, \frac{r_d}{d-1}, \ldots, \frac{r_d}{d-1}\right) \tag{22}$$

achieves equality $\mathrm{var}_\rho(-\log\rho) = N(d)$.

Theorem 8 is proved in Section 4.4 by the method of Lagrange multipliers. The proof shows that (22) is the unique ρ achieving equality, up to unitary equivalence.

Remark 9 (The quantities $N(d)$ and N_d)**.** $N(d)$ from Eq. (7) is well approximated by the easily computable $N_d \equiv \frac{1}{4}\log^2(d-1)+1$ since, by Lemma 14 (see also Fig. 1, lower right panel),

$$N_d \;>\; N(d) \;>\; N_d - 1\;. \tag{23}$$

One can even show $N(d) = N_d - O\left(1/\log^2 d\right)$ for $d \to \infty$, the optimal r in (7) being $r_d = 1/2 - 1/\log(d-1) + O\left(1/\log^2 d\right)$. Instead of the minimization (7), one may compute $N(d)$ numerically by finding the optimal $r = r_d \in [0, 1/2]$ as the (unique) solution of $(1-2r)\log\frac{1-r}{r}(d-1) = 2$ and plugging it back.

Note that the quantity $N(d)$ from the optimal upper bound (20) appears in the quadratic Taylor term of the optimal lower bound $M(\Delta, d)$ in (11) as well (cf. Remark 4). This can be understood in a pedestrian way by minimizing $D(\rho+\varepsilon\|\rho)$ at fixed ρ and for small ε (with $[\rho,\varepsilon] = 0$; see beginning of the proof of Theorem 1) under the constraint $S(\rho+\varepsilon) - S(\rho) = \delta$ (small), which gives $\delta^2/2\mathrm{var}_\rho(-\log\rho) + O(\delta^3)$. Finally minimizing this over all ρ, the quadratic term of $M(\delta, d)$ is therefore $\delta^2/2N(d)$ by Theorem 8.

2.2.2 Maximum heat capacity in finite dimensions

We now explain the thermodynamic significance of Theorem 8. Let H be a *Hamiltonian* of a d-dimensional system, i.e. a Hermitian $d \times d$-matrix (diagonal for classical systems). Then, at any *temperature* $T \in (0, \infty)$, the corresponding *thermal* (or equilibrium) state is

$$\rho_T \;:=\; \frac{e^{-H/T}}{\mathrm{tr}\left[e^{-H/T}\right]}\;, \tag{24}$$

with units chosen such that Boltzmann's constant $k_B = 1$. The (average) *energy* of the thermal state is $E(T) := \operatorname{tr}[H\rho_T]$, and the *heat capacity* $C(T)$ quantifies the rate of change of the system energy upon temperature variation:

$$C(T) := \left.\frac{dE}{dT}\right|_T = \frac{d}{dT}\operatorname{tr}\left[H\frac{e^{-H/T}}{\operatorname{tr}\left[e^{-H/T}\right]}\right] = \operatorname{var}_{\rho_T}(H/T) = \operatorname{var}_{\rho_T}(-\log\rho_T) , \tag{25}$$

where we omitted the little computation of the derivative, and used in the last step that the variance is unchanged under addition of a constant term (proportional to $\mathbb{1}$).

Eq. (25) shows that the heat capacity does not depend on H and T separately, but only on the thermal state ρ_T. Note that every full-rank state ρ can be interpreted as the thermal state of some Hamiltonian, and common extensions of the above framework include even some (or all) non-full-rank states; it is for example conventional to allow $T \in [0, \infty]$ and define ρ_0 to be the normalized projector onto the ground space of H, $H/\infty := 0$, and $C(\infty) := \lim_{T\to\infty} C(T)$.

Further note that, by (25), the heat capacity also equals the energy fluctuations $\operatorname{var}_{\rho_T}(H)$, i.e. the second moment of the energy, up to a factor of T^2.

Theorem 8 has thus the following corollary:

Corollary 10 (Maximum heat capacity in d dimensions)**.** *Let H be any Hamiltonian on a d-dimensional system, and let $T \in [0, \infty]$. Then its heat capacity $C(T)$ is uniformly bounded in terms of the dimension: For $d \geq 2$,*

$$C(T) \leq N(d) < N_d \equiv \frac{1}{4}\log^2(d-1) + 1 , \tag{26}$$

with $N(d)$ from Eq. (7). For $d = 1$, $C(T) = 0$.

Note that the first bound in (26) is tight for any d: The optimal state ρ from (22) has full-rank and is thus the thermal state of the Hamiltonian $H := -\log\rho$ at temperature $T := 1$.

Remark 11 (Exponential family of optimal states (10) and (22))**.** The optimal states ρ and σ from (10) come, for all values of Δ, from the same exponential family: defining a d-dimensional "Hamiltonian" $H_{opt} := \operatorname{diag}(-1, 0, \ldots, 0)$, it is $\sigma, \rho = e^{-H_{opt}/T_{\sigma,\rho}}/\operatorname{tr}\left[e^{-H_{opt}/T_{\sigma,\rho}}\right]$ for some "temperatures" $T_{\sigma,\rho} \in [0, \infty]$. The same is true for the state (22) having maximal surprisal variance or heat capacity; thermal states with one large occupation number (eigenvalue) $\approx 1/2$ and completely degenerate small occupations have thus the largest energy fluctuations [Mac03].

On an N-particle system, e.g. the space $\mathbb{C}^d = (\mathbb{C}^l)^{\otimes N}$ of N l-level particles, the Hamiltonian H_{opt} means physically that the system energy is minimized (-1) when each of the N particles is in a preferred state $|0\rangle$ and equals 0 otherwise, irrespective of the specific state. This very strong interaction between all N particles leads, at some temperature T_{crit}, to the largest possible heat capacity of any $d = l^N$-dimensional system by Corollary 10,

$$C(T_{crit}) = N(l^N) \simeq N_{d=l^N} \simeq \frac{1}{4}\log^2 l^N = N^2\,\frac{\log^2 l}{4} . \tag{27}$$

This is in stark contrast to a system of N independent (non-interacting) particles, whose heat capacity is proportional to N, i.e. "extensive", whereas (27) is faster than extensive.

When a system's heat capacity $C(T)/N$ per particle diverges at some temperature $T = T_{crit}$, one sometimes speaks of a second-order phase transition, and the system can then absorb or release energy density by just "reorganizing" its state without temperature change [Mac03]. Corollary 10 shows explicitly that such effects cannot occur for finite(-dimensional) systems.

3 Applications

Here we outline some implications for thermodynamics and information theory of Theorem 1, the inequality relating relative entropy and entropy difference (see Section 2.1).

3.1 Thermodynamics applications

In Section 3.1.1 we examine how slowly equilibration processes [AG12] have to be conducted to make them (close to) *thermodynamically reversible*. A relation between an intensive and an extensive quantity in many-particle systems is given in Section 3.1.2. Regarding the extensivity of the heat capacity in many-body systems, see also the previous Remark 11.

The following sections also serve to illustrate the prominence of relative entropy and entropy difference in thermodynamics and statistical physics.

3.1.1 Approach to reversibility in equilibration processes

In thermodynamics it is a common assumption (which can be justified in specific models) that a system with a Hamiltonian H and in weak interaction with an environment at temperature T will "equilibrate" to the thermal final state $\rho_f = e^{-H/T}/\mathrm{tr}\left[e^{-H/T}\right]$ (see Section 2.2.2), irrespective of its initial state ρ_i. The system's energy change associated with such a spontaneous state change is called *heat flow* or *heat* ΔQ [PW78, AG12]:

$$\Delta Q \;:=\; \mathrm{tr}\left[(\rho_f - \rho_i)H\right] \;=\; T\,\mathrm{tr}\left[(\rho_f - \rho_i)(-\log \rho_f)\right] \;. \tag{28}$$

One can relate this to the system's entropy change $\Delta S := \Delta(\rho_f, \rho_i) = S(\rho_f) - S(\rho_i)$:

$$\frac{\Delta Q}{T} \;=\; \Delta S - D(\rho_i \| \rho_f) \;\leq\; \Delta S \;. \tag{29}$$

(In order for all quantities to be well-defined, we assume ρ_f to be a full-rank state, i.e. assume $T \in (0, \infty]$; for simplicity and without further mentioning, we assume all states in this section to be of full-rank or at least of the same support.)

The above equilibration processes can also be conducted in a stepwise fashion, which was presented and analyzed in detail by Anders and Giovannetti [AG12]. One can view this as an attempt to formalize the vague notion of "slowness" of an equilibration process, which according to common physics folklore should make the process "thermodynamically reversible". We now recapitulate some elements from [AG12] and complement their analysis by a lower bound on how "close" a process can be to reversibility.

In a k-step process, adjust the system Hamiltonian successively first to H_1, then instantaneously to H_2, ..., and finally to $H_k \equiv H$, and let the system equilibrate with an environment at temperature T_j in each step $j = 1, \ldots, k$ (often, it will be either $H_j \equiv H$ for all j, or $T_j \equiv T$ for all j). We denote the associated intermediate thermal states by $\rho_j := e^{-H_j/T_j}/\mathrm{tr}\left[e^{-H_j/T_j}\right]$ (note, $\rho_k = \rho_f$) and define $\rho_0 := \rho_i$. The entropy change ΔS of the overall process equals just the sum of all changes $\Delta(\rho_j, \rho_{j-1})$, and the sum of the single-step quantities $\Delta Q_j/T_j$ satisfies, by (29),

$$\sum_{j=1}^{k} \frac{\Delta Q_j}{T_j} \;=\; \sum_{j=1}^{k} \left[\Delta(\rho_j, \rho_{j-1}) - D(\rho_{j-1} \| \rho_j)\right] \;=\; \Delta S - \sum_{j=1}^{k} D(\rho_{j-1} \| \rho_j) \;\leq\; \Delta S \;. \tag{30}$$

The inequality between the process quantity on the LHS and ΔS is the *Clausius Theorem* [AG12], often cited to be an incarnation of the Second Law of Thermodynamics. Note that, for $T_j \equiv T$, the LHS is just proportional to the total heat flow into the system, $\sum_j \Delta Q_j$.

In the special case [AG12] where intermediate steps $j = 1, \ldots, k-1$ are chosen such that the states ρ_j interpolate linearly between $\rho_i = \rho_0$ and $\rho_f = \rho_k$, i.e.

$$\rho_j \;=\; \left(1 - \frac{j}{k}\right) \rho_i + \frac{j}{k}\, \rho_f \qquad \text{for } j = 0, \ldots, k\,, \tag{31}$$

then the LHS of (30) can also be lower-bounded in terms of the entropy difference [AG12]:

$$\sum_{j=1}^{k} \frac{\Delta Q_j}{T_j} \;=\; \Delta S \,-\, \frac{D(\rho_f \| \rho_i) + D(\rho_i \| \rho_f)}{k} \,+\, \sum_{j=1}^{k} D(\rho_j \| \rho_{j-1}) \tag{32}$$

$$\geq\; \Delta S \,-\, \frac{D(\rho_f \| \rho_i) + D(\rho_i \| \rho_f)}{k}\;. \tag{33}$$

Thus, as the number of steps k in the interpolation (31) becomes finer (and if ρ_i, ρ_f have the same support), one has $\sum_j \Delta Q_j / T_j \to \Delta S$. This is remarkable since a priori the quantity $\sum_j \Delta Q_j / T_j$ depends on the details of the process, whereas $\Delta S = \Delta(\rho_f, \rho_i)$ depends only on its initial and final state.

Any process $\rho_i \mapsto \rho_1 \mapsto \ldots \mapsto \rho_f$ satisfying equality $\sum_j \Delta Q_j / T_j = \Delta S$ is called *(thermodynamically) reversible*, as intuitively one expects that the reverse of such a process leads back to the original situation. This intuition can be made rigorous for the process (31): The entropy production $\Delta S' = -\Delta S$ of the reverse process $\rho_f \mapsto \rho_{k-1} \mapsto \ldots \mapsto \rho_i$ exactly cancels ΔS, and also the process quantity $\sum_j \Delta Q'_j / T'_j$ will come close to $\Delta S' \approx -\sum_j \Delta Q_j / T_j$ by reasoning analogous to (30) and (33). For constant temperatures $T_j \equiv T$, the last fact means that (almost) no heat is produced during the entire cyclic process $\rho_i \mapsto \ldots \mapsto \rho_f \mapsto \ldots \mapsto \rho_i$, i.e. (almost) none of the work expended to (gradually) alter the Hamiltonian [PW78] is converted to heat, which physically is a less useful form of energy than work. In actual physical realizations, thermodynamic processes become irreversible when the system state $\rho(t)$ is not at all times t close to the thermal state determined by the system Hamiltonian $H(t)$ and the environment temperature $T(t)$. This happens for example when the process is conducted too fast so that the system cannot fully equilibrate at each infinitesimal step.

From this reasoning, one can quantify the *degree of irreversibility* of the process $\rho_i \mapsto \rho_1 \mapsto \ldots \mapsto \rho_f$ by the quantity $\sum_{j=1}^{k} D(\rho_{j-1} \| \rho_j)$ in (30). This corresponds to the amount of work wasted at least as heat in any cyclic completion $\rho_i \mapsto \rho_1 \mapsto \ldots \mapsto \rho_f = \rho_k \mapsto \rho_{k+1} \mapsto \ldots \mapsto \rho_{k+m} \equiv \rho_i$, since $\sum_{j=1}^{k+m} \Delta Q_j / T_j \leq -\sum_{j=1}^{k} D(\rho_{j-1} \| \rho_j)$ by (30). Quantitatively, denoting the minimal temperature $T_{min} := \min_{1 \leq j \leq k} T_j$, the excess heat production is at least

$$W_{waste} \;\geq\; T_{min} \sum_{j=1}^{k} D(\rho_{j-1} \| \rho_j)\;, \tag{34}$$

which is exact if $T_j \equiv T$ for all j. Theorem 1 now lower-bounds the sum in (34):

$$\sum_{j=1}^{k} D(\rho_{j-1} \| \rho_j) \;\geq\; \sum_{j=1}^{k} M\left(\Delta(\rho_{j-1}, \rho_j), d\right) \;=\; k \sum_{j=1}^{k} \frac{1}{k} M(\Delta(\rho_{j-1} \| \rho_j), d) \tag{35}$$

$$\geq\; k\, M\left(\sum_{j=1}^{k} \frac{1}{k} \Delta(\rho_{j-1}, \rho_j),\, d\right) \;=\; k\, M\left(\frac{-\Delta S}{k},\, d\right) \tag{36}$$

$$\geq\; \frac{1}{k}\, \frac{(\Delta S)^2}{3 \log^2 d}\;, \tag{37}$$

where $d < \infty$ denotes the dimension of the system, the second inequality is by convexity of the function M (Theorem 2), and we exemplarily used the lower bound (12).

Achieving a degree ε of reversibility by a stepwise process thus necessitates a minimum number $k = O(1/\varepsilon)$ of steps via Eq. (37). When k interpreted as the time duration of the entire process – assuming that each equilibration step consumes roughly equal time – then (37) substantiates the folklore whereby thermodynamically reversible processes have to be conducted "infinitely slowly". Our estimate is thus relevant for fundamental thermodynamics and especially for small systems [SBL$^+$11], as it delineates where the idealized but commonplace notion of *reversible process* can apply. It also provides new heat bounds for processes out of equilibrium in the area of non-equilibrium thermodynamics [Lin83, Jar99, Jar11].

Although the lower bound (36) is essentially tight in the typical thermodynamics situation where only the entropy difference ΔS between two states is known, it becomes trivial for $\Delta S = 0$. In this and other cases, when in addition the initial and final states ρ_i, ρ_f are known, one may use an estimate similar to (35)–(37) but based on Pinsker's inequality (17):

$$\sum_{j=1}^{k} D(\rho_{j-1}\|\rho_j) \;\geq\; \sum_{j=1}^{k} \frac{1}{2}\|\rho_{j-1}-\rho_j\|_1^2 \;\geq\; \frac{k}{2}\left(\sum_{j=1}^{k}\frac{1}{k}\|\rho_{j-1}-\rho_j\|_1\right)^2 \;\geq\; \frac{\|\rho_i-\rho_f\|_1^2}{2k}\,. \tag{38}$$

On the topic of stepwise processes we finally remark that the approach to reversibility $\sum_{j=1}^{k}\Delta Q_j/T_j \to \Delta S$ for $k\to\infty$ is *not* special to the linear interpolation process (31) [AG12]. Rather, for any (piecewise continuously differentiable) curve $\rho(t)$ in state space with $\rho(0)=\rho_i$, $\rho(1)=\rho_f$, a discretization at points $0=t_0<t_1<\ldots<t_k=1$ gives

$$\sum_{j=1}^{k}\frac{\Delta Q_j}{T_j} \;=\; \sum_{j=1}^{k}\mathrm{tr}\left[(-\log\rho(t_j))(\rho(t_j)-\rho(t_{j-1}))\right] \tag{39}$$

$$\to\; \int_{t=0}^{1}\mathrm{tr}\left[(-\log\rho(t))\,d\rho(t)\right] \;=\; \int_0^1 dt\,\mathrm{tr}\left[-\dot\rho(t)\log\rho(t)\right] \tag{40}$$

$$=\; \int_0^1 dt\,\frac{d}{dt}\mathrm{tr}\left[\rho(t)-\rho(t)\log\rho(t)\right] \;=\; S(\rho_f)-S(\rho_i) \;=\; \Delta S\;, \tag{41}$$

with convergence as the discretization becomes finer, $k\to\infty$ and $\max_j|t_j-t_{j-1}|\to 0$ (i.e. a Riemann sum). Thus, any state change $\rho_i\mapsto\rho_f$ can be made thermodynamically reversible (when $\mathrm{supp}[\rho_i]=\mathrm{supp}[\rho_f]$). For the discretized process $\rho(t)$ we do however not have a lower convergence estimate as in (33) (the upper bound from the Clausius Theorem (30) holds of course for any discretization).

In this section, we have considered thermalizing processes, bringing an arbitrary state ρ_i to a thermal state ρ_f, and have measured the heat production w.r.t. the Hamiltonian H corresonding to the *final* (thermal) state [AG12]. This leads to the Clausius inequality (30).

In [RW13] we use Theorem 1 in the reverse situation where an initially thermal state ρ_i is used as the resource in a process leading away from equilibrium. The heat production is again measured w.r.t. the system's Hamiltonian, which there however is related to the *initial* state and reverses the inequality (30) [AG12, RW13]. Furthermore, the paper [RW13] concerns a bipartite scenario – the Landauer process involving a system and a thermal reservoir [Lan61] – where a Second Law-like statement can be formulated more properly and where the above stepwise process may be implemented by swapping the system and reservoir states.

3.1.2 Free energy vs. entropy density

To further elucidate the thermodynamic meaning of the quantity $D(\rho_i\|\rho_f)$ for a thermal final state $\rho_f = e^{-H/T}/\mathrm{tr}\left[e^{-H/T}\right]$ (cf. Eq. (29) in Section 3.1.1), we relate it to the *work extractable at constant temperature* from the state ρ_i, and then examine it in a many-particle system.

For this, consider an *isothermal process*, i.e. where the temperature T remains constant and only the Hamiltonian is changed from its initial value $H_0 \equiv H$ in k successive steps to H_1, ..., $H_k \equiv H$, at each of which the system equilibrates as in Section 3.1.1. The total heat flow $\Delta Q := \sum_{j=1}^{k} \Delta Q_j$ during the process then satisfies the Clausius inequality $T\Delta S \geq \Delta Q$ (Eq. (30)), so that

$$\begin{aligned} T\, D(\rho_i\|\rho_f) &= -\operatorname{tr}\left[H(\rho_f - \rho_i)\right] + T\left[S(\rho_f) - S(\rho_i)\right] = F(\rho_i) - F(\rho_f) && (42) \\ &\geq -\Delta E + \Delta Q = -\Delta W\ , && (43) \end{aligned}$$

where we have defined: the *free energy* $F(\rho) := \operatorname{tr}[H\rho] - TS(\rho)$ of a state ρ (at temperature T and for Hamiltonian H), the *internal energy increase* $\Delta E := \operatorname{tr}[H(\rho_f - \rho_i)]$, and the *work* $\Delta W := \Delta E - \Delta Q$ done on the system [PW78, AG12].

According to Section 3.1.1, equality in (43) can be approached by a suitable (reversible) process (note that then the jump from $H = H_0$ to the first equilibration step $H_1 \approx -T \log \rho_i$ may be big, whereas the further steps $H_1 \mapsto \ldots \mapsto H_k = H$ are small). Thus, the amount $T\, D(\rho_i\|\rho_f) = (-\Delta W)_{max}$ of work *can* be extracted from the state ρ_i by a thermodynamic process at temperature T and using the internal energy function H. Conversely, for given temperature T and Hamiltonian H, this is also the *maximum* amount of work extractable from ρ_i since, for any process leading to a final state ρ'_f (not necessarily thermal for either H or T),

$$\begin{aligned} -\Delta W' &= -\Delta E' + \Delta Q' \leq -\operatorname{tr}\left[H(\rho'_f - \rho_i)\right] + T\left[S(\rho'_f) - S(\rho_i)\right] \\ &= F(\rho_i) - F(\rho'_f) = \left[F(\rho_i) - F(\rho_f)\right] - \left[F(\rho'_f) - F(\rho_f)\right] \\ &= T\, D(\rho_i\|\rho_f) - T\, D(\rho'_f\|\rho_f) \leq T\, D(\rho_i\|\rho_f)\ , \end{aligned} \tag{44}$$

where the last inequality amounts to the "thermodynamic inequality" [OP93], i.e. the fact that the free energy $F(\rho)$ attains its minimum at the thermal state $\rho = \rho_f$ (uniquely for $T \neq 0$).

Theorems 1 and 2 thus lower-bound the extractable work at constant temperature T:

$$W_{extr,T} = T\, D(\rho_i\|\rho_f) \geq T\, M(-\Delta S, d) \gtrsim 2T \left(\frac{\Delta S}{\log d}\right)^2 = 2T \left(\frac{\Delta s}{\log l}\right)^2\ , \tag{45}$$

where in the last step we have assumed a system of N l-level particles, i.e. $d = l^N$ (cf. Remark 11), and defined the change in *entropy density* $\Delta s := \Delta S/N = (S(\rho_f) - S(\rho_i))/N$ [BR97].

Inequality (45) seems quite unusual as its LHS is the "extensive" free energy difference or extractable work whereas the RHS is an "intensive" quantity, given by the entropy density and temperature; moreover, in the "thermodynamic limit" $N \to \infty$, the inequality is essentially tight. The reason for this is that states attaining equality are of the form (10), which are strongly correlated as discussed in Remark 11, such that one cannot speak of few-particle properties and the designation "extensive" is not appropriate.

3.2 Information-theoretic applications

We have already outlined in Section 2.2.1 the meaning of $\sqrt{\operatorname{var}_\rho(\log \rho)}$ as the fluctuation in codeword length of an optimal prefix code; for a source with d distinct signals this fluctuation is at most $\simeq \frac{1}{2} \log d$ by Theorem 8. In the following, we discuss implications in information theory of the lower bound on the relative entropy (Theorems 1 and 2).

3.2.1 Cost of wrong code, universal codes, and Shannon channel capacity

For a source producing i.i.d. signals i according to a classical probability distribution $\rho = \{p_i\}_{i=1}^{d}$, Shannon's source compression theorem [Sha48, CT06] shows any prefix code with a D-ary alphabet to have an average length of at least $S(\rho)/\log D$ per encoded signal. The lower the

entropy of the signal distribution, the shorter on average the encoded message can be. This length is in fact achievable – up to less than 1 alphabet symbol – by assigning codewords of length $\lceil -\log_D p_i \rceil$ to the signals.

If one however wrongly assumes the signals i to be distributed according to $\sigma = \{q_i\}_{i=1}^d$ and constructs a code for this distribution, with codewords of length $\lceil -\log_D q_i \rceil$, then the average code length L_σ will be

$$L_\sigma \;=\; \sum_{i=1}^d p_i \lceil -\log_D q_i \rceil \;\geq\; -\frac{1}{\log D} \sum_{i=1}^d p_i \log q_i \;=\; \frac{S(\rho)}{\log D} + \frac{D(\rho\|\sigma)}{\log D} \;. \tag{46}$$

The last term is the cost of the wrong code [CT06] beyond the optimal average code length $S(\rho)/\log D$ when one knew the correct distribution.

Theorems 1 and 2 give a lower bound on this penalty just in terms of the difference $\delta = (S(\rho) - S(\sigma))/\log D$ between the supposed length $S(\sigma)/\log D$ and the optimal achievable length $S(\rho)/\log D$:

$$\frac{D(\rho\|\sigma)}{\log D} \;\geq\; \frac{M(\delta \log D, d)}{\log D} \;\geq\; \delta^2 \frac{2\log D}{\log^2(d-1)+4} \;, \tag{47}$$

where the last inequality holds only for positive expected savings $\delta \geq 0$.

When the signals $i \in \{1, \ldots, d\}$ follow one of the distributions ρ^θ, where the parameter $\theta \in \{1, \ldots, m\}$ is not known, one may choose a coding distribution σ ("universal code") that minimizes the maximal occuring penalty or *redundancy* (see [CT06], Section 13.1):

$$R^* \;:=\; \min_\sigma \max_\theta D(\rho^\theta\|\sigma) \;. \tag{48}$$

The theorems from Section 2.1 give an easily computable lower bound on the quantity R^*: For this, denote by S_{min} and S_{max} be the minimal resp. maximal entropy $S(\rho^\theta)$ among the states ρ^θ. Then using the properties from Theorem 2 and Remark 5, we have:

$$\begin{aligned}
R^* \;&\geq\; \min_\sigma \max_\theta\; M(S(\rho^\theta) - S(\sigma), d) \;=\; \min_{S\in[0,\log d]} \max_\theta\; M(S(\rho^\theta) - S, d) && (49)\\
&=\; \min_{S\in[S_{min},S_{max}]} \max\left\{ M\left(S_{max} - S, d\right),\, M\left(S_{min} - S, d\right) \right\} && (50)\\
&\geq\; \min_{S\in[S_{min},S_{max}]} \max\left\{ M\left(S - S_{max}, d\right),\, M\left(S_{min} - S, d\right) \right\} && (51)\\
&=\; M\left(\frac{1}{2}(S_{min} - S_{max}),\, d\right) \;, && (52)
\end{aligned}$$

where in the third line we *assumed* $M(|\Delta|, d) \geq M(-|\Delta|, d)$, which seems true numerically and is consistent with all previous analytical results, but which we have not proved; as it at least holds for the lower bounds (11), lower-bounding (52) with the help of (11) gives true statements.

The minimal redundancy R^* from (48) equals the Shannon capacity $C(T)$ (measured in nats) of the classical discrete memoryless channel $T : \theta \mapsto i$ that is defined by the transition probabilities $T(i|\theta) := p_i^\theta$ (Theorem 13.1.1 in [CT06], originally due to [Gal79, Rya79]; the proof uses a minimax theorem). This gives, with the same provisions as below Eq. (52):

Proposition 12 (Lower bound on the classical Shannon capacity)**.** *For a discrete memoryless channel $T : \mathcal{X} \to \mathcal{Y}$, given by transition probabilities $T(y|x)$ and with finite output dimension $|\mathcal{Y}| \geq 2$, the Shannon capacity $C(T)$ is lower-bounded as*

$$C(T) \;\geq\; M\left(-\frac{S_{max} - S_{min}}{2},\, |\mathcal{Y}|\right) \;, \tag{53}$$

where S_{max} and S_{min} denote the maximal and minimal entropies, respectively, of any column $T(\cdot|x)$ of the transition matrix. Here, M denotes the function defined in Eq. (6), which in turn can be lower-bounded as in Theorem 2.

The RHS in Eq. (53), or one of its lower bounds from (11), is easier to evaluate than Shannon's mutual information formula for the exact $C(T)$ [Sha48, CT06]. Lower-bounding the relative entropies in (48) by Pinsker's or the H-O-T inequality (17) [Csi67, HOT81, AE05], would lead to a linear program in the variables σ. Eq. (53) also provides a more systematic way to obtain lower bounds on $C(T)$ than by plugging trial input distributions into Shannon's formula.

On the other hand, the lower bound (53) will be trivial iff all columns $T(\cdot|x)$ have the same entropy, whereas the capacity $C(T)$ vanishes only iff all columns are themselves identical. Also, the lower bound in (53) can never exceed $(\log 2) = 1$ bit, since it has to hold for input dimension $|\mathcal{X}| = 2$ as well (or when there are only two distinct columns in $T(\cdot|x)$); in the most favorable case $S_{max} - S_{min} = \log d$, the RHS of (53) is actually always between $0.111 \simeq 0.16$ bit (for $d = 2$) and $\log \sqrt{3} \simeq 0.80$ bit (for $d \to \infty$; cf. Remark 4).

In the quantum setting, identical formulas apply for the cost of the wrong code (47) and the redundancy (52), see [SW00, SW01]. Furthermore, the Holevo quantity, which is a lower bound on the classical capacity of a quantum channel [NC00], equals the relative entropy radius of the channel output, i.e. the redundancy (48) over all output states [OPW97, SW00]. For a quantum channel, however, there is no systematic way known in particular to find the *minimum output entropy* S_{min} efficiently; the channel output set has, e.g., generally infinitely many extreme points.

3.2.2 Hypothesis testing and large deviations

The relative entropy features prominently also in hypothesis testing and large deviation theory [CT06]. On the one hand, relative entropies $D(\sigma\|\rho)$ between given states σ, ρ appear for example as error exponents in asymmetric hypothesis testing (in the classical Chernoff-Stein Lemma [CT06] as well as in its quantum analogue [HP91, ON00]), such that Theorems 1 and 2 apply immediately to yield lower bounds on error decay rate in terms of the entropy difference $S(\sigma) - S(\rho)$ only.

On the other hand, in these areas one is often interested in quantities like

$$\operatorname{dist}(E, \rho) \;:=\; \inf_{\sigma \in E} D(\sigma\|\rho) \;, \tag{54}$$

where E is some set of d-dimensional probability distributions and ρ a fixed distribution. Sometimes the set E is described by an entropy constraint, for example in universal coding for all d-dimensional sources of entropy less than R ([CT06]; similarly [BDK$^+$05] for the quantum case): here, the decoding error probability vanishes exponentially in the message length n like $\sim \exp(-n \operatorname{dist}(E, \rho))$ if the true source distribution is ρ (assuming $S(\rho) < R$) and where $E := \{\sigma | S(\sigma) > R\}$. The decay rate, $\operatorname{dist}(E, \rho)$, may thus be lower-bounded by $M(R - S(\rho), d)$ according to Theorems 1 and 2 simply in terms of an entropy difference.

Finally, in symmetric hypothesis testing between two classical (commuting) probability distributions ρ_1, ρ_2, the optimal error decay rate is given by the *Chernoff information* $\xi(\rho_1, \rho_2) = -\log \min_{0 \le s \le 1} \operatorname{tr}\left[\rho_1^s \rho_2^{1-s}\right]$ [Che52, CT06], which has the property that there exists a distribution σ (from the Hellinger arc between ρ_1 and ρ_2) satisfying $\xi(\rho_1, \rho_2) = D(\sigma\|\rho_1) = D(\sigma\|\rho_2)$. Similar to the derivation leading up to (52), the latter quantity can be lower-bounded in terms of the entropy difference $\Delta(\rho_1, \rho_2) = S(\rho_1) - S(\rho_2)$ between the two states only:

$$\xi(\rho_1, \rho_2) \;\ge\; \frac{|\Delta(\rho_1, \rho_2)|^2}{2\left(\log^2(d-1) + 4\right)} \;-\; \frac{|\Delta(\rho_1, \rho_2)|^3}{3\left(\log^2(d-1) + 4\right)^2} \;, \tag{55}$$

where the last expression does not involve any extremization (cf. Theorem 2).

Whereas for symmetric hypothesis testing between (non-commuting) quantum states ρ_1, ρ_2 the basic formula for the decay rate $\xi(\rho_1, \rho_2)$ holds as well, the existence of a state σ as above is not known [ANS+08]. We can therefore not apply the same reasoning to get a lower bound on $\xi(\rho_1, \rho_2)$ in the quantum setting. For other kinds of (dimension-independent) bounds on the quantum and classical Chernoff information, see [ANS+08, Aud12].

3.2.3 Mutual information

Let ρ_{AB} be a joint state on a bipartite system AB with respective local dimensions d_A and d_B and total dimension $d = d_A d_B$ (in the classical probabilistic case, ρ_{AB} is a joint probability distribution of two random variables A and B with d_A and d_B outcomes, respectively). Then its *mutual information* $I(A:B) := S(\rho_A) + S(\rho_A) - S(\rho_{AB})$ can be written as both a relative entropy and an entropy difference [OP93]:

$$I(A:B) = S(\rho_A \otimes \rho_B) - S(\rho_{AB}) = -\Delta(\rho_{AB}, \rho_A \otimes \rho_B) \tag{56}$$

$$= D(\rho_{AB} \| \rho_A \otimes \rho_B) \ , \tag{57}$$

where ρ_A, ρ_B denote the reduced states (marginal probability distributions) for A and B, and in the first line we used the notation (2).

Here we just remark that Theorem 1, which relates relative entropy and entropy difference, does not give any constraints in this situation: For $\Delta \in [-\log d, 0]$, which is the case here, it is $-\Delta \geq M(\Delta, d)$ by Remark 5, with strict inequality except for $\Delta = -\log d, 0$; the fact $D(\rho_{AB} \| \rho_A \otimes \rho_B) = -\Delta(\rho_{AB} \| \rho_A \otimes \rho_B)$ is thus consistent with Theorem 1 and therefore (9) does not give new information.

Note that $I(A:B) \leq \min\{\log d_A, \log d_B\}$ in the classical case, whereas for quantum states $I(A:B) \leq 2\min\{\log d_A, \log d_B\}$, so that the maximum value $\log d = \log d_A + \log d_B$ of $-\Delta(\rho_{AB}, \rho_A \otimes \rho_B)$ and of $D(\rho_{AB} \| \rho_A \otimes \rho_B)$ can be attained only in the quantum case and only when $d_A = d_B$ with a maximally entangled state ρ_{AB} [NC00].

4 Proofs

4.1 Proof of Theorem 1

Proof of Theorem 1. To prove the inequality (9) and the optimality statement around (10), we will compute, for any fixed $\Delta \in [-\log d, \log d]$, the infimum

$$\inf_{\sigma,\rho} \left\{ D(\sigma\|\rho) \;\middle|\; S(\sigma) - S(\rho) = \Delta \right\} \tag{58}$$

over d-dimensional quantum states σ, ρ, and show that it equals $M(\Delta, d)$ from Eq. (6) with optimal states σ, ρ of the form (10).

We first note that the infimum in (58) is attained: For $\Delta = \log d$, one necessarily has $\sigma = \mathbb{1}/d$ and ρ is a pure state, so that $D(\sigma\|\rho) = \infty$ is "attained"; on the other hand, this equals $M(\log d, d) = \infty$, as $\Delta = \log d$ in (6) enforces $s = (d-1)/d$ and $r = 0$; the case $\Delta = \log d$ is thus done and we exclude it from all further considerations. For any $\Delta \in [-\log d, \log d)$, there exists a full-rank state ρ with $S(\sigma) - S(\rho) = \Delta$, and thus the intersection of the set of all pairs (σ, ρ) satisfying $S(\sigma) - S(\rho) = \Delta$ with the set of all pairs satisfying $\mathrm{supp}[\sigma] \subseteq \mathrm{supp}[\rho]$ is non-empty and compact, so that the infimum of the continuous function $D(\sigma\|\rho)$ over this set is attained and finite. For similar reasons, the infimum in (6) is attained. For this reasoning and for the argumentation below, we note that $H(s) + s\log(d-1)$ is strictly increasing in $s \in [0, (d-1)/d]$

from the value 0 at $s = 0$ to $\log d$ at $s = (d-1)/d$ with first derivative

$$\frac{d}{ds}\left(H(s) + s\log(d-1)\right) \;=\; \log\frac{1-s}{s}(d-1) \qquad \text{for } s \in (0,1)\ . \tag{59}$$

It is easy to see that the infimum in (58) is attained for *commuting* states σ and ρ: Fixing the state ρ and fixing all eigenvalues $\mathrm{spec}(\sigma)$ of σ (which also fixes the entropy $S(\sigma)$; this should be done to be consistent with $S(\sigma) - S(\rho) = \Delta$ for the fixed Δ), the infimum (over σ) of the relative entropy

$$D(\sigma\|\rho) \;=\; -S(\sigma) + \mathrm{tr}\left[(-\log\rho)\sigma\right] \tag{60}$$

is attained by the state σ which is diagonal in the same basis as $(-\log\rho)$ and has its eigenvalues ordered in the opposite way as $(-\log\rho)$ [Bha97]; as the logarithm is a strictly increasing function, σ will thus also be diagonal in the same basis as ρ (and in particular commute with ρ), with its eigenvalues ordered in the same way as ρ. (When $\mathrm{rank}(\rho) < \mathrm{rank}(\mathrm{spec}[\sigma])$, the infimum is $+\infty$, and this as well can be attained by a σ commuting with ρ). This commutativity carries over to the infimum in (58), and implies that the bound we are about to prove will be optimal for the case of classical d-dimensional probability distributions (i.e. diagonal density matrices) as well.

One can get more information about the optimal pair (σ, ρ) from Klein's inequality, i.e. the non-negativity of the relative entropy. We fix again the state ρ and fix the entropy of σ to equal $S(\sigma) = S$, leaving the spectrum of σ otherwise free; under these constraints we again minimize (60). In thermodynamics language (see Eq. (24) and below), this is the minimization of the "energy" of σ w.r.t. the "Hamiltonian" $(-\log\rho)$ under the entropy constraint $S(\sigma) = S$; by the "thermodynamic inequality", a version of Klein's inequality, it is well-known that the minimum is attained for a "thermal state" $\sigma \sim e^{-\gamma(-\log\rho)}$, i.e. $\sigma = \rho^\gamma/\mathrm{tr}\left[\rho^\gamma\right]$, for some "inverse temperature" $\gamma \in [0, +\infty]$ (here we define $0^0 := 0$, and $\rho^\infty/\mathrm{tr}\left[\rho^\infty\right]$ is to be understood as the maximally mixed state on the eigenspace of ρ corresponding to its largest eigenvalue).

Making this argument more precise requires some care: We consider the minimization of (60) under variation of *both* σ and ρ with the constraints of fixed $S(\rho) = S_\rho$ and fixed $S(\sigma) = S_\sigma = S_\rho + \Delta$, and denote by $(\widehat{\sigma}, \widehat{\rho})$ a minimizing assignment. Only in the case $S_\rho = 0$ can we have $D(\widehat{\sigma}\|\widehat{\rho}) = +\infty$, and we do not consider this case here as it is only necessary if $\Delta = \log d$, which was already discussed above. Thus $D(\widehat{\sigma}\|\widehat{\rho}) < \infty$, and so we have $\mathrm{supp}[\widehat{\sigma}] \subseteq \mathrm{supp}[\widehat{\rho}]$, which implies $\log\mathrm{rank}(\widehat{\rho}) \geq S_\sigma$. Now, if $S_\sigma = \log\mathrm{rank}(\widehat{\rho})$, then obviously $\widehat{\sigma} = \widehat{\rho}^0/\mathrm{tr}\left[\widehat{\rho}^0\right]$ (i.e. $\widehat{\sigma}$ is the maximally mixed state on the support of $\widehat{\rho}$; we define $0^0 := 0$). Second, if $\log\mathrm{rank}(\widehat{\rho}) > S_\sigma > \log m_0$, where m_0 denotes the dimension of the eigenspace of the largest eigenvalue of $\widehat{\rho}$ (i.e. the dimension of the ground state space of the "Hamiltonian" $(-\log\widehat{\rho})$), then due to continuity of the entropy [Fan73] there exists $\gamma \in (0, \infty)$ with $S(\widehat{\rho}^\gamma/\mathrm{tr}\left[\widehat{\rho}^\gamma\right]) = S_\sigma$. We claim that then $\widehat{\sigma} = \widehat{\rho}^\gamma/\mathrm{tr}\left[\widehat{\rho}^\gamma\right]$ is the unique minimizer of (60) under variation of σ (when keeping $\rho = \widehat{\rho}$ fixed). This is easy to see by verifying $\gamma\left(D(\tau\|\widehat{\rho}) - D(\widehat{\rho}^\gamma/\mathrm{tr}\left[\widehat{\rho}^\gamma\right]\|\widehat{\rho})\right) = D(\tau\|\widehat{\rho}^\gamma/\mathrm{tr}\left[\widehat{\rho}^\gamma\right])$ for all states τ with $S(\tau) = S_\sigma$, and then using that $D(\tau\|\widehat{\rho}^\gamma/\mathrm{tr}\left[\widehat{\rho}^\gamma\right]) \geq 0$ with equality iff $\tau = \widehat{\rho}^\gamma/\mathrm{tr}\left[\widehat{\rho}^\gamma\right]$ (by Klein's inequality). Third, if $S_\sigma = \log m_0$, then the maximally mixed state on the eigenspace of the largest eigenvalue of $\widehat{\rho}$ is obviously the unique state with entropy S_σ and minimizing (60), i.e. we could formally write $\widehat{\sigma} = \widehat{\rho}^\infty/\mathrm{tr}\left[\widehat{\rho}^\infty\right]$. Fourth, if $S_\sigma < \log m_0$, then $\widehat{\sigma}$ may be any state supported on the eigenspace of the largest eigenvalue of $\widehat{\rho}$. In all of these case, $\widehat{\sigma}$ and $\widehat{\rho}$ commute, which was already seen before by simpler reasoning.

For the following we can thus assume that

$$\widehat{\sigma} \;=\; \mathrm{diag}(\widehat{q}_1, \ldots, \widehat{q}_d) \qquad \text{and} \qquad \widehat{\rho} \;=\; \mathrm{diag}(\widehat{p}_1, \ldots, \widehat{p}_d)\ . \tag{61}$$

Now fixing $\sigma = \widehat{\sigma}$ in (60), the minimization over all commuting states $\rho = \mathrm{diag}(p_1, \ldots, p_d)$ leads to the Lagrange function

$$L(\{p_i\}, \nu, \mu) \;:=\; \sum_i (\widehat{q}_i \log \widehat{q}_i - \widehat{q}_i \log p_i) \;+\; \nu \sum_i p_i \;+\; \mu \sum_i p_i \log p_i \tag{62}$$

with Lagrange multipliers ν and μ corresponding to the nomalization and entropy constraints $\mathrm{tr}\,[\rho] = 1$ and $S(\rho) = S_\rho$, respectively. We now look at this as a function of those variables p_i, for which the corresponding element $\widehat{p}_i \neq 0$ is positive (i.e. which lie in the interior of the domain of L), and we fix the other elements p_i to be zero. Then, since $p_i = \widehat{p}_i$ is a minimizing assignment, by the method of Lagrange multipliers one is guaranteed the existence of $\widehat{\nu}$, $\widehat{\mu} \in (-\infty, +\infty)$ such that

$$\left.\frac{dL}{dp_j}\right|_{\{\widehat{p}_i\},\widehat{\nu},\widehat{\mu}} \;=\; -\,\frac{\widehat{q}_j}{\widehat{p}_j} \;+\; (\widehat{\nu} + \widehat{\mu}) \;+\; \widehat{\mu} \log \widehat{p}_j \;=\; 0 \qquad \forall j \;\text{ with }\; \widehat{p}_j \neq 0\;. \tag{63}$$

This excludes that the fourth case from the previous paragraph can be a minimizing case, since in this case there are $\widehat{p}_j = \widehat{p}_k = \lambda_{max}(\widehat{\rho}) > 0$ and $\widehat{q}_j \neq \widehat{q}_k$, contradicting (63). Within the third case of the previous paragraph, it excludes the possibility that, apart from the maximum eigenvalue $\lambda_{max}(\widehat{\rho})$, there could be two further distinct non-zero eigenvalues $\widehat{p}_i \neq \widehat{p}_j$, as in the third case both of these would have corresponding $\widehat{q}_i = \widehat{q}_j = 0$, again contradicting (63). Thus, in the third case above, $\widehat{\rho}$ has at most two distinct non-zero eigenvalues, as does $\widehat{\sigma}$.

Also for the first and second cases of the above paragraph we now want to show that, except possibly when $\gamma = 1$ (i.e. for $\widehat{\sigma} = \widehat{\rho}$ or $\Delta = 0$), $\widehat{\rho}$ has at most two distinct non-zero eigenvalues, and $\widehat{\sigma}$ as well. In these two cases, we have $\widehat{q}_j = \widehat{p}_j^{\,\gamma}/Z$ for some $\gamma \in [0, \infty)$ with $Z := \sum_i \widehat{p}_i^{\,\gamma} > 0$. Now define $x_j := Z\widehat{q}_j/\widehat{p}_j = \widehat{p}_j^{\,\gamma-1}$ for each j with $\widehat{p}_j > 0$. Eq. (63) says then that, for $\gamma \neq 1$, the points x_j lie at intersections of the non-horizontal affine function $-x/Z + (\widehat{\nu} + \widehat{\mu})$ with the function $-(\widehat{\mu}/(\gamma - 1)) \log x$ (both are functions of $x > 0$). The latter function is either strictly convex or strictly concave or constant (depending on whether the prefactor is negative or positive or zero). The two functions can thus not intersect at more than 2 distinct points $x_j > 0$. When $\gamma \neq 1$, there can therefore be at most 2 distinct non-zero values of x_j, i.e. also at most 2 distinct non-zero values of $\widehat{p}_j$ and of $\widehat{q}_j$.

Summing up so far, any states ρ and σ attaining the infimum in (58) commute and, for $\Delta \neq 0$, have at most two distinct non-zero eigenvalues each, in such a way that distinct eigenvalues in σ and in ρ correspond to each other. More precisely,

$$\begin{aligned} \sigma \;&= \mathrm{diag}\left(\frac{1-s}{m}, \ldots, \frac{1-s}{m}, \frac{s}{n}, \ldots, \frac{s}{n}, 0, \ldots, 0\right)\;, \\ \rho \;&= \mathrm{diag}\left(\frac{1-r}{m}, \ldots, \frac{1-r}{m}, \frac{r}{n}, \ldots, \frac{r}{n}, 0, \ldots, 0\right)\;, \end{aligned} \tag{64}$$

where $m, n \geq 1$, $m + n \leq d$ and $s, r \in [0, 1]$. Permuting the entries of both states simultaneously, we may assume the entries of σ to be ordered non-increasingly, i.e. $(1-s)/m \geq s/n$. The above analysis showed further that the diagonal entries of a minimizing pair are ordered in the same order (see below Eq. (60); this also can be seen by the fact the the inverse temperature γ above turned out to be always non-negative). Thus, $(1-r)/m \geq r/n$ as well, and we will therefore in the following always assume $0 \leq s, r \leq n/(m+n)$. Even in the case $\Delta = 0$, some of the minimizing pairs (σ, ρ) have this form (choose any m, n, and $s = r$), and we thus assume this form below; similarly for the case $\Delta = \log d$, where e.g. $m = 1$, $n = d - 1$, $s = (d-1)/d$, $r = 0$. We can thus continue the optimization in (58) with states of the form (64). Before that, note

for the states in (64):

$$S(\sigma) \;=\; H(s) + (1-s)\log m + s\log n \;, \quad S(\rho) \;=\; H(r) + (1-r)\log m + r\log n \;, \tag{65}$$

$$\Delta(\sigma,\rho) \;=\; S(\sigma) - S(\rho) \;=\; H(s) - H(r) + (s-r)\log\frac{n}{m} \;, \tag{66}$$

$$D(\sigma\|\rho) \;=\; D_2(s\|r) \;=\; s\log\frac{s}{r} + (1-s)\log\frac{1-s}{1-r} \;. \tag{67}$$

Given $\Delta \neq 0$, let now the states σ and ρ in (64), parametrized by s, r, m, and n, attain the infimum in (58). Our next goal is to show $m = 1$ and $n = d-1$. For now, we will denote by $\tau_{t,m,n}$ the state parametrized by t, m, and n, such that, for example, $\tau_{s,m,n} = \sigma$ and $\tau_{r,m,n} = \rho$ in (64). Assume that there exist $m', n' \geq 1$ with $m' + n' \leq d$ and $n'/m' > n/m$. We will then show that there exists some s', such that the pair of states $(\tau_{s',m',n'}, \tau_{r,m',n'})$ would achieve a strictly lower value in (58) than the pair (σ, ρ). For this, compute

$$S(\tau_{s,m',n'}) - S(\tau_{r,m',n'}) \;=\; H(s) - H(r) + (s-r)\log\frac{n'}{m'} \;=\; \Delta + (s-r)\log\frac{n'/m'}{n/m} \;, \tag{68}$$

and note the the last logarithm is positive due to $n'/m' > n/m$. Now, assume first $\Delta > 0$. Then, from (66), we have $s > r$ due to our convention $s, r \leq n/(m+n)$. Thus the expression (68) is strictly larger than Δ, and because its left-hand-side is an increasing function of the argument $s \leq n/(m+n)$ (similar to the computation (59)), there exists due to continuity some $s' \in (r, s)$ with

$$S(\tau_{s',m',n'}) - S(\tau_{r,m',n'}) \;=\; \Delta \;. \tag{69}$$

Since $D_2(s\|r)$ is strictly increasing in is first argument for $s \geq r$ (assuming $r > 0$, which holds due to $\Delta < \log d$), we have $D(s'\|r) < D(s\|r)$, which contradicts the optimality of the pair (σ, ρ). The case $\Delta < 0$ is analogous. We have thus shown that, if we choose the parametrization of the optimal pair in (64) such that $s \leq n/(m+n)$, then there do *not* exist $m', n' \geq 1$ with $m' + n' \leq d$ and $n'/m' > n/m$. This implies $n = d-1$, $m = 1$ for the optimal pair (σ, ρ).

Using now $n = d-1$, $m = 1$ in (64) and recalling (65)–(67), the optimal states (for $\Delta \neq 0$) will thus be of the form (10), where (s, r) attains the minimum in (6); for $\Delta = 0$, the optimal states can be chosen to be of that form. The preceding proof shows also that, for $\Delta \neq 0$, the optimal states are necessarily of the form (10), up to simultaneous unitary transformations of σ and ρ; the proof in Section 4.2 shows furthermore that, for each $\Delta \neq 0$, the optimal s and r are unique. For $\Delta = 0$, the optimal pairs are obviously exactly the ones with $\sigma = \rho$. □

4.2 Proof of Theorem 2

Proof of Theorem 2. $M(\Delta, d) \geq 0$ is clear, and the stated values are argued below Eq. (5).

For $N = N(d)$, the first inequality in (11) is just Lemma 13, and for $N \geq N(d)$ it follows from the monotonicity of the lower bound:

$$\frac{d}{dN}\left(Ne^{\frac{\Delta}{N}} - N - \Delta\right) \;=\; -e^{\frac{\Delta}{N}}\left[e^{-\frac{\Delta}{N}} - \left(1 - \frac{\Delta}{N}\right)\right] \;\leq\; 0 \;, \tag{70}$$

since the square brackets is non-negative due to convexity of the exponential function. For any N and Δ, the second inequality in (11) is easily verified by subtracting both sides from each other and observing that the difference and its first three derivatives w.r.t. Δ vanish at $\Delta = 0$, whereas the fourth derivative is positive everywhere. If one defines, as usual, the minimum over

an empty set in (6) to be ∞, then the lower bounds (11) hold even for Δ outside the range $[-\log d, \log d]$.

To prove (12), we use the left inequality in (11) and find a constant $c > 0$ satisfying $N(d)e^{\Delta/N(d)} - N(d) - \Delta \geq \Delta^2/c\log^2 d$ for all $d \geq 0$ and $\Delta \in [-\log d, \log d]$. At fixed d, this inequality holds for all Δ iff it holds for $\Delta = -\log d$. For large $d \to \infty$, one sees by Taylor expansion and by $N(d) = N_d + O(1) = (\log^2 d)/4 + O(1)$ that any $c > 1/2$ is sufficient. Examining small $d \geq 2$ numerically, one sees that $c = 3$ is sufficient.

For the convenient upper bounds on N_d, see Lemma 14.

We now sketch a proof of strict convexity (and continuous differentiability) of $M(\Delta, d)$, which is somewhat involved. For this, we employ its definition (6), will somtimes abbreviate $D := \log(d-1) \geq 0$, and denote by r_d the (unique) $r \in (0, 1/2)$ attaining the maximum in (7), i.e. satisfying $(1-2r_d)\log\frac{1-r_d}{r_d}(d-1) = 2$. We also define $\gamma_d \in (0, (d-1)/d)$ to be the unique solution of $(1-\gamma_d)\log\frac{1-\gamma_d}{\gamma_d}(d-1) = 1$; one can check $\gamma_d > r_d$.

If, for some $\Delta = x \in (-\log d, \log d)$, a pair $(s, r) \in (0, (d-1)/d)^2$ attains the minimum in (6), then by the method of Lagrange multipliers the following two equations hold:

$$\Delta(s,r) \;:= H(s) - H(r) + (s-r)D \;=\; x\;, \tag{71}$$

$$F(s,r) \;:= \left(\log\frac{1-r}{r} - \log\frac{1-s}{s}\right)\left(D + \log\frac{1-r}{r}\right) - \left(\frac{s}{r} - \frac{1-s}{1-r}\right)\left(D + \log\frac{1-s}{s}\right) \;=\; 0\;, \tag{72}$$

where the latter equality expresses the requirement that the gradients of the target function and the constraint function be parallel (i.e., that the 2×2-matrix formed by these gradients have vanishing determinant). In a small enough neighborhood of any such pair $(s, r) \in (0, (d-1)/d)^2$ with $s \neq r$, the equations (71)–(72) are sufficiently well-behaved to have a unique solution $(s(x'), r(x'))$ for any $x' \in (x - \varepsilon, x + \varepsilon)$, as the solution of the differential equations obtained from (71)–(72). For any $s = r$, (71)–(72) are satisfied with $x = 0$ (corresponding to the trivial optimality cases $\sigma = \rho$), but near any such point there are no other pairs with $F(s, r) = 0$ and $s \neq r$ (as one sees from a quadratic expansion of $F(s, r)$) with the exception of $s = r = r_d$: around $x = 0$ and $s = r = r_d$, the equations (71)–(72) have a solution with $\dot{s}(x = 0) = (1 - 2r_d)/3$, $\dot{r}(x = 0) = -(1 - 2r_d)/6$ (overdots denote derivatives w.r.t. x), which can be seen by computing third directional derivatives of $F(s, r)$ at this point.

Examining the equation $F(s, r) = 0$ for $(s, r) \in (0, (d-1)/d)^2$ (by way of discussing $F(s, r)$ and its derivative $F_s(s, r)$ along each fixed r) and furthermore considering optimal pairs (s, r) for any $\Delta = x$ in (6) on the boundary of $[0, (d-1)/d]^2$, one finds the following: For $r = 0$, optimal pairs are obtained for $s = 0$ and for $s = (d-1)/d$ (where $x = \log d$); for $0 < r < r_d$, optimal pairs are obtained for $s = r$ and for one other value $s \in (r_d, (d-1)/d)$ (where $0 < x < \log d$); for $r = r_d$, the only optimal pair is obtained for $s = r_d$ (where $x = 0$); for $r_d < r < \gamma_d$, optimal pairs are obtained for one value $s \in (0, r_d)$ (where $x \in (\Delta_r, 0)$, where we define $\Delta_r := \Delta(s = 0, r = \gamma_d) = 1 - A + \log\gamma_d \in (-\log d, 1 - \log d)$) and for $s = r$; for $\gamma_d \leq r \leq (d-1)/d$, optimal pairs are obtained for $s = 0$ (where $x \in [-\log d, \Delta_r]$) and for $s = r$.

Combining this with the above differentiability result and defining $s(0) := r(0) := r_d$ for $x = 0$ while disregarding the other optimal pairs with $s = r$, we get the following: For any $x \in [-\log d, \log d] \setminus \{0\}$ there exists exactly one optimal pair $(s(x), r(x))$ (i.e. with $\Delta(s(x), r(x)) = x$), the curve $(s(x), r(x))$ is continuous in $x \in [-\log d, \log d]$, and differentiable in $x \in (\Delta_r, \log d)$. Thus already, $M(x, d) = D(s(x)\|r(x))$ is continuous in $x \in [-\log d, \log d]$ (with the usual convention $\lim_{x \nearrow \log d} M(x, d) = \infty = M(\log d, d)$).

We can now finally prove strict convexity of $M(x, d)$. First, for $x \in [-\log d, \Delta_r]$, it is $s(x) = 0$. One can thus explicitly write $\Delta = -H(r) - Dr$ as a function of $M = M(x, d) = D_2(s = 0\|r) =$

$-\log(1-r)$ in this range of $\Delta = x$; the function $\Delta = \Delta(M)$ is easily seen to be continuously differentiable, strictly decreasing and strictly convex in this range. Its inverse $M = M(\Delta, d)$ is thus strictly convex as well and continuously differentiable in $\Delta \in (-\log d, \Delta_r]$, and one can compute $dM/d\Delta\,|_{\Delta=\Delta_r} = -1$ (and $dM/d\Delta\,|_{\Delta\searrow -\log d} = -\infty$).

Second, for $x \in (\Delta_r, \log d)$, the optimal pairs $(s(x), r(x)) \in (0, (d-1)/d)^2$ satisfy (71)–(72). We can thus compute

$$\begin{aligned}
\frac{d}{dx} M(x,d) &= \frac{d}{dx} D_2(s(x)\|r(x)) && (73)\\
&= \left(\log\frac{1-r(x)}{r(x)} - \log\frac{1-s(x)}{s(x)}\right)\dot{s}(x) - \left(\frac{s(x)}{r(x)} - \frac{1-s(x)}{1-r(x)}\right)\dot{r}(x) && (74)\\
&= \left(\log\frac{1-r(x)}{r(x)} - \log\frac{1-s(x)}{s(x)}\right)\left(D + \log\frac{1-s(x)}{s(x)}\right)^{-1}, && (75)
\end{aligned}$$

where in the last step we used (72) and the derivative of (71) w.r.t. x. Notice for later that $dM(x,d)/dx\,|_{x\searrow\Delta_r} = -1$ since $s(x) \searrow 0$ and $r(x) \to \gamma_r$ for $x \searrow \Delta_r$. Thus,

$$\begin{aligned}
\left(D + \log\frac{1-s(x)}{s(x)}\right)^2 \frac{d^2}{dx^2} M(x,d) &= \left(D + \log\frac{1-r(x)}{r(x)}\right)\frac{\dot{s}(x)}{s(x)(1-s(x))}\\
&\quad - \left(D + \log\frac{1-s(x)}{s(x)}\right)\frac{\dot{r}(x)}{r(x)(1-r(x))}\,. \qquad (76)
\end{aligned}$$

Strict convexity, $d^2M(x,d)/dx^2 > 0$, would thus follow from $\dot{s}(x) \geq 0$ and $\dot{r}(x) \leq 0$; to see the last implication, note that not both of $\dot{s}(x)$ and $\dot{r}(x)$ can vanish simultaneously because of $d\Delta(s(x), r(x))/dx = 1 > 0$. The last insight also shows that $\dot{s}(x) \leq 0$ and $\dot{r}(x) \geq 0$ cannot both be true simultaneously unless $\dot{s}(x) = \dot{r}(x) = 0$. It thus suffices now to show that $\dot{s}(x)$ and $\dot{r}(x)$ cannot both be simultaneously positive nor both be simultaneously negative. For $x = 0$, this was remarked above. For $x \in (\Delta_r, \log d) \setminus \{0\}$, we show it in the following way.

Differentiating (72), one has

$$0 = \frac{d}{dx} F(s(x), r(x)) = F_s(s(x), r(x))\,\dot{s}(x) + F_r(s(x), r(x))\,\dot{r}(x)\,. \qquad (77)$$

The considerations of the equation $F(s,r) = 0$ above show that $F_s(s(x), r(x)) > 0$ for $s(x) \neq r(x)$. Finally, the fact that $s(x) > r(x)$ implies $r(x) < r_d$ and the fact that $s(x) < r(x)$ implies $r(x) > r_d$ (see above) can be used, together with (72), to show $F_r(s(x), r(x)) > 0$ for $s(x) \neq r(x)$. (77) then implies that not both of $\dot{s}(x)$ and $\dot{r}(x)$ can have the same sign.

$M(x,d)$ is thus strictly convex in $x \in (\Delta_r, \log d)$, as well as in $x \in [-\log d, \Delta_r]$. Since $M(x,d)$ is continuous with matching left-sided and right-sided derivatives at $x = \Delta_r$ (see above), it is strictly convex in the whole range $x \in [-\log d, \log d]$. Continuity of $(s(x), r(x))$ and Eq. (75), together with the above considerations of the range $x \in [-\log d, \Delta_r]$, finally prove continuous differentiability of $M(x,d)$ in $x \in (-\log d, \log d)$.

□

4.3 Auxiliary Lemmas

Lemma 13 (Simple lower bound on $M(\Delta, d)$)**.** *For $2 \leq d < \infty$ and $\Delta \in [-\log d, \log d]$, the quantity $M(\Delta, d)$ from Eq. (6) is lower-bounded as follows:*

$$M(\Delta, d) \geq N(d)\left(e^{\frac{\Delta}{N(d)}} - 1 - \frac{\Delta}{N(d)}\right), \qquad (78)$$

where $N(d)$ is defined in Eq. (7).

Proof. Define the function $\Delta(s,r) := H(s) - H(r) + (s-r)\log(d-1)$. To show Lemma 13, we will prove

$$G(s,r) \;:=\; D_2(s\|r) \;-\; N(d)\left(e^{\frac{\Delta(s,r)}{N(d)}} - 1 - \frac{\Delta(s,r)}{N(d)}\right) \;\geq\; 0 \tag{79}$$

for all $s, r \in [0, (d-1)/d]$. The statement is easily verified for $r = 0$, since $D_2(s\|0) = +\infty$ unless $s = 0$. We thus fix $r \in (0, (d-1)/d]$ from now on, so that $G(s,r)$ is a function of $s \in [0, (d-1)/d]$.

At $s = r$, the function $G(s = r, r) = 0$ vanishes, as does its first derivative

$$\frac{d}{ds}G(s,r)\bigg|_{s=r} \;=\; \log\frac{1-r}{r} - \log\frac{1-s}{s} - \left(e^{\frac{\Delta(s,r)}{N(d)}} - 1\right)\log\frac{1-s}{s}(d-1)\bigg|_{s=r} \;=\; 0\;. \tag{80}$$

Furthermore, $G(s,r)$ is convex in $s \in [0, (d-1)/d]$ since, for $s \in (0, (d-1)/d]$,

$$\frac{d^2}{ds^2}G(s,r) \;=\; e^{\frac{\Delta(s,r)}{N(d)}}\,\frac{1}{N(d)\,s(1-s)}\left[N(d) \;-\; s(1-s)\left(\log\frac{1-r}{r}(d-1)\right)^2\right] \;\geq\; 0 \tag{81}$$

as the term in square brackets is non-negative due to the definition of $N(d)$ in Eq. (7).

All of this together shows that, for each fixed $r \in [0, (d-1)/d]$, $G(s,r)$ attains its minimum 0 at $s = r$, which finally proves (79). □

Lemma 14 (Simple bounds on $N(d)$)**.** *For $d \geq 2$, the optimization $N(d)$ from Eq. (7) is upper-bounded in the following ways:*

$$N_d - 1 \;=\; \frac{1}{4}\log^2(d-1) \;<\; N(d) \;<\; N_d \;=\; \frac{1}{4}\log^2(d-1) + 1\;, \tag{82}$$

$$N(d) \;<\; \log^2 d\;, \tag{83}$$

where N_d in the first inequality was defined in Eq. (8).

Proof. To prove the upper bound in (82), we show that for all $r \in [0,1]$,

$$0 \;<\; \frac{1}{4}\log^2(d-1) + 1 \;-\; r(1-r)\left(\log\frac{1-r}{r}(d-1)\right)^2\;. \tag{84}$$

For $r = 0, 1$ this is clear due to the convention $0 \cdot \infty = 0$ (or by continuity), and for $r = 1/2$ it is easily verified. Let thus $r \in (0,1) \setminus \{1/2\}$. The right-hand-side of (84) equals

$$\begin{aligned}
&= \left(\frac{1}{2} - r\right)^2\log^2(d-1) - 2r(1-r)\left(\log\frac{1-r}{r}\right)\log(d-1) + 1 - r(1-r)\left(\log\frac{1-r}{r}\right)^2 \\
&= \left(\left(\frac{1}{2} - r\right)\log(d-1) - \frac{r(1-r)}{\frac{1}{2}-r}\right)^2 + 1 - r(1-r)\left(\log\frac{1-r}{r}\right)^2 - \frac{r^2(1-r)^2}{\left(\frac{1}{2}-r\right)^2}\left(\log\frac{1-r}{r}\right)^2 \\
&\geq \frac{1}{(1-2r)^2}\left[(1-2r)^2 - r(1-r)\left(\log\frac{1-r}{r}\right)^2\right]\;,
\end{aligned} \tag{85}$$

where the inequality arises by omitting the non-negative first term $(\ldots)^2$ from the step before.

Now, the last expression does not depend on the dimension d anymore, and one can show that it is positive for all $r \in (0,1) \setminus \{1/2\}$. This is numerically easily verified, or analytically in the following way: The term in square brackets in (85) vanishes at $r = 1/2$, as do its first three derivatives w.r.t. r, whereas its fourth derivative

$$\frac{d^4}{dr^4}\Big[\text{from (85)}\Big] \;=\; \frac{8}{r^2(1-r)^2} + \frac{2(1-2r)^2}{r^3(1-r)^3} + (1-2r)\left(\log\frac{1-r}{r}\right)\frac{16r(1-r) + 4(1-2r)^2}{r^3(1-r)^3}$$

is strictly positive for all $r \in (0,1)$, since $(1-2r)\log\frac{1-r}{r} \geq 0$ for $r \in (0,1)$.

The lower bound in (82) follows by letting $r \to 1/2$ in the definition (7) of $N(d)$.

To prove (83), we look for a constant c satisfying $N(d) \leq c\log^2 d$. For large $d \to \infty$ any $c > 1/4$ works by (82). Examining small $d \geq 2$ numerically, one finds that $c \geq 0.92$ suffices. □

4.4 Proof of Theorem 8

Proof of Theorem 8. For fixed $d \geq 2$, we maximize the expression on the LHS of (20) or (19) over all probability distributions $\{p_i\}$ (i.e., spectra of ρ), which leads to the Lagrange function

$$L(\{p_i\},\nu) \;:=\; \sum_i p_i(\log p_i)^2 - \left(\sum_i p_i \log p_i\right)^2 + \nu\sum_i p_i \;, \tag{86}$$

with the Lagrange multiplier ν corresponding to the normalization $\mathrm{tr}\,[\rho] = 1$. Assume now that $\{\widehat{p}_i\}$ (corresponding to the state $\widehat{\rho}$) attains the maximum of (19) over all probability distributions $\{p_i\}$ (due to continuity and compactness, this maximum is attained). We now view (86) as a function of those variables p_i for which $\widehat{p}_i > 0$, fixing the other elements p_i to be zero. Then, due to the extremality of $\{\widehat{p}_i\}$ and having components in the interior of the domain of L, the method of Lagrange multipliers guarantees the existence of $\widehat{\nu} \in (-\infty,+\infty)$ such that

$$\begin{aligned} 0 \;&=\; \left.\frac{dL}{dp_j}\right|_{\{\widehat{p}_i\},\widehat{\nu}} \;=\; (\log\widehat{p}_j)^2 + 2\log\widehat{p}_j - 2\left(\sum_i \widehat{p}_i\log\widehat{p}_i\right)(1+\log\widehat{p}_j) + \widehat{\nu} \\ &=\; (S(\{\widehat{p}_i\}) + 1 + \log\widehat{p}_j)^2 \;-\; (S(\{\widehat{p}_i\}))^2 \;+\; \widehat{\nu} - 1 \qquad\quad \forall j \;\text{ with }\; \widehat{p}_j > 0 \;, \end{aligned} \tag{87}$$

where the quantity $S(\{\widehat{p}_i\}) = S(\widehat{\rho})$ denotes the entropy of the distribution $\{\widehat{p}_i\}$ and in particular does not depend on the index j. Thus, the equality (87) implies that

$$\log\widehat{p}_j \;=\; \pm\sqrt{(S(\widehat{\rho}))^2 - \widehat{\nu} + 1} \;-\; S(\widehat{\rho}) - 1 \qquad\quad \forall j \;\text{ with }\; \widehat{p}_j > 0 \;, \tag{88}$$

so that strict monotonicity of the logarithm yields that there can be at most two distinct non-zero elements in $\{\widehat{p}_i\}$.

Thus, leaving off hats again, an optimal $\rho = \widehat{\rho}$ has the form

$$\rho \;=\; \left(\frac{1-r}{m},\ldots,\frac{1-r}{m},\frac{r}{n},\ldots,\frac{r}{n},0,\ldots,0\right) \tag{89}$$

with $m,n \geq 1$, $m+n \leq d$, $r \in [0,1]$. W.o.l.g. we can assume $r \leq 1/2$ by permuting the entries of ρ. For such states one has, after a small calculation,

$$\mathrm{var}_\rho(\log\rho) \;=\; r(1-r)\left(\log\frac{1-r}{r} + \log\frac{n}{m}\right)^2 \;. \tag{90}$$

Maximizing this, for any fixed $r \in [0,1/2]$, over m and n yields $n = d-1$ and $m = 1$. Maximizing (90) finally over r gives a unique $r = r_d \in (0,1/2)$, namely unique the value of $r \in [0,1/2]$ satisfying $(1-2r)\log\frac{1-r}{r}(d-1) = 2$, and the maximum of (90) is $N(d)$ from (7).

The inequality (21) is shown by Lemma 14, which completes the proof of Theorem 8. □

Acknowledgments. We thank Daniel Reitzner and Marco Tomamichel for helpful discussion. DR was supported by the Marie Curie Intra European Fellowship QUINTYL. MMW acknowledges support from the Alfried Krupp von Bohlen und Halbach-Stiftung.

5 References

[Abe11] J. Aberg, "Truly work-like work extraction", arXiv:1110.6121 [quant-ph] (2011).

[AG12] J. Anders, V. Giovannetti, "Thermodynamics of discrete quantum processes", arXiv:1211.0183 [quant-ph] (2012).

[Aud07] K. M. R. Audenaert, "A sharp continuity estimate for the von Neumann entropy", J. Phys. A 40, 8127-8136 (2007).

[Aud12] K. M. R. Audenaert, "Quantum sandwich bounds", arXiv:1207.1197 [quant-ph] (2012).

[AE05] K. M. R. Audenaert, J. Eisert, "Continuity bounds on the quantum relative entropy", J. Math. Phys. 46, 102104 (2005).

[ANS+08] K. M. R. Audenaert, M. Nussbaum, A. Szkola, F. Verstraete, "Asymptotic error rates in quantum hypothesis testing", Comm. Math. Phys. 279, 251-283 (2008).

[Bha97] R. Bhatia, "Matrix Analysis", *Springer*, Heidelberg (1997).

[BDK+05] I. Bjelakovic, J. D. Deuschel, T. Krüger, R. Seiler, Ra. Siegmund-Schultze, A. Szkola, "A quantum version of Sanov's theorem", Comm. Math. Phys. 260, 659-671 (2005).

[BR97] O. Bratteli, D. W. Robinson, "Operator Algebras and Quantum Statistical Mechanics 2", 2nd. ed., *Springer*, Berlin (1997).

[Che52] H. Chernoff, "A Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", Ann. Math. Stat. 23, 493-507 (1952).

[CT06] T. M. Cover, J. A. Thomas, "Elements of Information Theory", 2nd. ed., *Wiley-Interscience*, Hoboken (2006).

[Csi67] I. Csiszar, "Information type measure of difference of probability distributions and indirect observations", Studia Sci. Math. Hungar. 2, 299-318 (1967).

[EDR+12] D. Egloff, O. C. O. Dahlsten, R. Renner, V. Vedral, "Laws of thermodynamics beyond the von Neumann regime", arXiv:1207.0434 [quant-ph] (2012).

[Fan73] M. Fannes, "A continuity property of the entropy density for spin lattice systems", Commun. Math. Phys. 31, 291-294 (1973).

[Gal79] R. G. Gallager, "Source coding with side information and universal coding", Tech. Rept. LIDS-P-937, Laboratory for Information Decision Systems, MIT, Cambridge, MA (1979).

[GMM10] J. Gemmer, M. Michael, G. Mahler, "Quantum Thermodynamics", 2nd. ed., *Springer*, Berlin (2010).

[HOT81] F. Hiai, M. Ohya, M. Tsukuda, "Sufficiency, KMS condition and relative entropy in von Neumann algebras", Pacific J. Math. 96, 99-109 (1981).

[HP91] F. Hiai, D. Petz, "The proper formula for relative entropy and its asymptotics in quantum probability", Commun. Math. Phys. 143, 99-114 (1991).

[Jar99] C. Jarzynski, "Microscopic analysis of Clausius-Duhem processes", J. Stat. Phys. 96, 415 (1999).

[Jar11] C. Jarzynski, "Equalities and inequalities: Irreversibility and the Second Law of Thermodynamics at the nanoscale", Annu. Rev. Condens. Matter Phys. 2, 329-351 (2011).

[KL51] S. Kullback, R. A. Leibler, "On information and sufficiency", Ann. Math. Statist. 22, 79-86 (1951).

[Lan61] R. Landauer, "Irreversibility and heat generation in the computing process", IBM J. Res. Dev. 5, 183 (1961).

[Li12] K. Li, "Second order asymptotics for quantum hypothesis testing", arXiv:1208.1400 [quant-ph] (2012).

[Lin83] G. Lindblad, Non-equilibrium entropy and irreversiblity, D. Reidel Publishing Company, Dordrecht (1983).

[Mac03] D. J. C. MacKay, "Information Theory, Inference, and Learning Algorithms", *Cambridge University Press*, Cambridge (2003).

[NC00] M. A. Nielsen, I. L. Chuang, "Quantum Computation and Quantum Information", *Cambridge University Press*, Cambridge (2000).

[ON00] T. Ogawa, H. Nagaoka, "Strong converse and Stein's lemma in quantum hypothesis testing", IEEE Trans. Inf. Theory 46, 2428 (2000).

[OP93] M. Ohya, D. Petz, "Quantum entropy and its use", *Springer*, Berlin (1993).

[OPW97] M. Ohya, D. Petz, N. Wanatabe, "On capacities of quantum channels", Prob. Math. Stat. 17, 179-196 (1997).

[PPV10] Y. Polyanskiy, H. V. Poor, S. Verdu, "Channel coding rate in the finite blocklength regime", IEEE Trans. Inf. Theory 56, 2307-2359 (2010).

[PW78] W. Pusz, S. L. Woronowicz, "Passive states and KMS states for general quantum systems", Comm. Math. Phys. 58, 273-290 (1978).

[RW13] D. Reeb, M. M. Wolf, "General microscopic proof of Landauer's Principle and finite-size improvements", paper accepted for a talk at TQC 2013 (University of Guelph, Canada), extended arXiv submission in preparation (2013).

[Ren05] R. Renner, "Security of Quantum Key Distribution", Ph.D. thesis, ETH Zürich (2005); see also arXiv:quant-ph/0512258.

[Rya79] B. Y. Ryabko, "Encoding of a source with unknown but ordered probabilities", Probl. Inf. Transm., 134-138 (1979).

[Sha48] C. E. Shannon, "A mathematical theory of communication", Bell Syst. Tech. J. 27, 379-423 (1948).

[Sch95] B. Schumacher, "Quantum coding", Phys. Rev. A 51, 2738-2747 (1995).

[SW00] B. Schumacher, M. D. Westmoreland, "Relative entropy in quantum information theory", in: "Quantum Computation and Quantum Information: A Millenium Volume", S. Lomonaco (ed.), AMS Contemporary Mathematics series [arXiv:quant-ph/0004045] (2000).

[SW01] B. Schumacher, M. D. Westmoreland, "Indeterminate-length quantum coding", Phys. Rev. A 64, 042304 (2001).

[SBL+11] P. Skrzypczyk, N. Brunner, N. Linden, S. Popescu, "The smallest refrigerators can reach maximal efficiency", J. Phys. A: Math. Theor. 44, 492002 (2011).

[TH12] M. Tomamichel, M. Hayashi, "A hierarchy of information quantities for finite block length analysis of quantum tasks", arXiv:1208.1478 [quant-ph] (2012).

[Ume62] H. Umegaki, "Conditional expectation in an operator algebra, IV (entropy and information)", Kodai Math. Sem. Rep. 14, 59-85 (1962).

[vN32] J. von Neumann, "Mathematische Grundlagen der Quantenmechanik", *Springer*, Berlin (1932); in English: "Mathematical Foundations of Quantum Mechanics", translated by Robert T. Beyer, *Princeton University Press* (1955).

[Weh78] A. Wehrl, "General properties of entropy", Rev. Mod. Phys. 50, 221-260 (1978).